APPLIED OPTIMIZATION

Joseph Geunes
Elif Akçali
Panos M. Pardalos
H. Edwin Romeijn
Zuo-Jun (Max) Shen

APPLICATIONS OF SUPPLY CHAIN MANAGEMENT AND E-COMMERCE RESEARCH

Springer

Applications of Supply Chain Management and E-Commerce Research

Applied Optimization

Volume 92

Applications of Supply Chain Management and E-Commerce Research

Edited by

JOSEPH GEUNES
University of Florida, Gainesville, U.S.A.

ELIF AKÇALI
University of Florida, Gainesville, U.S.A.

PANOS M. PARDALOS
University of Florida, Gainesville, U.S.A..

H. EDWIN ROMEIJN
University of Florida, Gainesville, U.S.A.

ZUO-JUN (MAX) SHEN
University of Florida, Gainesville, U.S.A.

Springer

eBook ISBN: 0-387-23392-X
Print ISBN: 0-387-23391-1

Contents

Part I Supply Chain Operations

1
Coordination of Inventory and Shipment Consolidation Decisions: A Review of Premises, Models, and Justification 3
Sıla Çetinkaya

2
A Near-Optimal Order-Based Inventory Allocation Rule in an Assemble-to-Order System and its Applications to Resource Allocation Problems 53
Yalçın Akçay, Susan H. Xu

3
Improving Supply Chain Performance through Buyer Collaboration 87
Paul M. Griffin, Pınar Keskinocak, and Seçil Savaşaneril

4
The Impact of New Supply Chain Management Practices on the Decision Tools Required by the Trucking Industry 119
Jacques Roy

5
Managing the Supply-Side Risks in Supply Chains: Taxonomies, Processes, and Examples of Decision-Making Modeling 141
Amy Z. Zeng, Paul D. Berger, Arthur Gerstenfeld

6
Demand Propagation in ERP Integrated Assembly Supply Chains: Theoretical Models and Empirical Results 161
S. David Wu, Mary J. Meixell

Part II Electronic Commerce and Markets

7
Bridging the Trust Gap in Electronic Markets: A Strategic Framework for Empirical Study 195
Gary E. Bolton, Elena Katok, Axel Ockenfels

8
Strategies and Challenges of Internet Grocery Retailing Logistics 217
Tom Hays, Pınar Keskinocak, Virginia Malcome de López

9
Enabling Supply-Chain Coordination: Leveraging Legacy Sources for Rich Decision Support 253
Joachim Hammer, William O'Brien

10
Collaboration Technologies for Supporting E-supply Chain Management 299
Stanley Y. W. Su, Herman Lam, Rakesh Lodha, Sherman Bai, Zuo-Jun (Max) Shen

Part III From Research to Practice

11
The State of Practice in Supply-Chain Management: A Research Perspective 325
Leroy B. Schwarz

12
Myths and Reality of Supply Chain Management: Implications for Industry-University Relationships 363
André Kuper, Sarbani Bublu Thakur-Weigold

13
Supply Chain Management: Interlinking Multiple Research Streams 383
James C. Hershauer, Kenneth D. Walsh, Iris D. Tommelein

14
PROFIT: Decision Technology for Supply Chain Management at IBM Microelectronics Division 411
Ken Fordyce, Gerald (Gary) Sullivan

15
Case Studies: Supply Chain Optimization Models in a Chemical Company 453
Young M. Lee, E. Jack Chen

Foreword

In February 2002, the Industrial and Systems Engineering (ISE) Department at the University of Florida hosted a *National Science Foundation Workshop on Collaboration and Negotiation in Supply Chain Management and E-Commerce.* This workshop focused on characterizing the challenges facing leading-edge firms in supply chain management and electronic commerce, and identifying research opportunities for developing new technological and decision support capabilities sought by industry. The audience included practitioners in the areas of supply chain management and E-Commerce, as well as academic researchers working in these areas. The workshop provided a unique setting that has facilitated ongoing dialog between academic researchers and industry practitioners.

This book codifies many of the important themes and issues around which the workshop discussions centered. The editors of this book, all faculty members in the ISE Department at the University of Florida, also served as the workshop's coordinators. In addition to workshop participants, we also invited contributions from leading academics and practitioners who were not able to attend. As a result, the chapters herein represent a collection of research contributions, monographs, and case studies from a variety of disciplines and viewpoints. On the academic side alone, chapter authors include faculty members in supply chain and operations management, marketing, industrial engineering, economics, computer science, civil and environmental engineering, and building construction departments. Thus, throughout the book we see a range of perspectives on supply chain management and electronic commerce, both of which often mean different things to different disciplines. The subjects of the chapters range from operations research based models of supply chain planning problems to statements and perspectives on research and practice in the field. Three main themes serve to divide the book into three separate parts.

Part I of the book contains six chapters broadly focused on operations and logistics planning issues and problems. The first chapter, *Coordi-*

nation of Inventory and Shipment Consolidation Decisions: A Review of Premises, Models, and Justification, by Sıla Çentinkaya, provides a detailed and insightful look into the interaction between outbound logistics consolidation policies and inventory costs. This work focuses on providing both insights and guidance on effective policies for coordinating inventory and logistics decisions. Yalçin Akçay and Susan Xu study the component allocation problem in an assemble-to-order manufacturing environment in Chapter 2, *A Near-Optimal Order-Based Inventory Allocation Rule in an Assemble-to-Order System and its Applications to Resource Allocation Problems.* The problem is modeled as a multi-dimensional knapsack problem, and they develop an efficient heuristic for finding high-quality solutions to this problem. Their results provide insights on how to effectively manage assemble-to-order systems.

In Chapter 3, *Improving Supply Chain Performance through Buyer Collaboration,* Paul M. Griffin, Pınar Keskinocak, and Seçil Savaşaneril take a look at how different buyers can leverage collective purchase volumes to reduce procurement costs through collaboration. In addition to discussing recent trends in electronic markets and systems for procurement, the authors provide some very interesting results on the value of collaboration in procurement, both internally (across different divisions in the same organization) and externally (among different firms). In Chapter 4 *The Impact of New Supply Chain Management Practices on the Decision Tools Required by the Trucking Industry,* Jacques Roy provides an overview of the recent advances in supply chain management and information technologies, and discusses how the emerging information technologies can be used to support decision making to improve the efficiency of the freight transportation industry.

Chapter 5, *Managing the Supply-Side Risks in Supply Chains: Taxonomies, Processes, and Examples of Decision-Making Modeling,* by Amy Zeng, Paul Berger, and Arthur Gerstenfeld, analyzes the risks associated with suppliers and the supply market from a quantitative point of view. Two optimization-based decision tree models are proposed in order to answer questions of how many suppliers should be used and whether to use standby suppliers. In Chapter 6, *Demand Propagation in ERP Integrated Assembly Supply Chains: Theoretical Models and Empirical Results,* David Wu and Mary Meixell study supply chain demand propagation in an ERP-integrated manufacturing environment. They examine key factors that influence demand variance in the assembly supply chain, assess their effects, and develop insight into the underlying supply process.

Part II contains four chapters on electronic markets and E-Commerce technologies and their role in facilitating supply chain coordination.

Chapter 7, *Bridging the Trust Gap in Electronic Markets: A Strategic Framework for Empirical Study,* by Gary Bolton, Elena Katok, and Axel Ockenfels, describes a strategic framework for evaluating automated *reputation systems* for electronic markets, and provides suggestions on how to improve automated reputation system performance. In Chapter 8 *Strategies and Challenges of Internet Grocery Retailing Logistics,* Tom Hays, Pınar Keskinocak, and Virginia Malcome de López provide a detailed and thorough look at the practice of the Internet grocery retailing, focusing on alternative business models, order fulfillment and delivery methods. They offer a discussion of the lessons learned from failure and success stories of e-grocers, a summary of current trends, and future opportunities and directions.

Chapter 9, entitled *Enabling Supply-Chain Coordination: Leveraging Legacy Sources for Rich Decision Support,* by Joachim Hammer and William O'Brien, describes how firms with different legacy systems can use new technologies to not only reduce the cost of establishing inter-system communication and information sharing, but also to provide coordinated decision support in supply chains. The focus on information technologies for supporting effective supply chain management continues in Chapter 10, *Collaboration Technologies for Supporting E-supply Chain Management* (by Stanley Su, Herman Lam, Rakesh Lodha, Sherman Bai, and Max Shen). This chapter describes an e-supply chain management information infrastructure model to manage and respond to important supply chain "events" and to automate negotiation between channel members.

Part III provides a link between research and practice, beginning with three chapters that provide different frameworks, viewpoints, and paradigms on research and practice perspectives on supply chain management. The last two chapters illustrate industrial examples of effective application of supply chain management research in practice.

In Chapter 11, *The State of Practice in Supply-Chain Management: A Research Perspective,* Leroy Schwarz develops a new paradigm for managing supply chains, providing insight into the evolution of supply chain practice to date. From this perspective, he describes examples of current state-of-the-art practice in supply chain management, and forecasts future practice. In Chapter 12 *Myths and Reality of Supply Chain Management: Implications for Industry- University Relationships,* André Kuper and Sarbani Bublu Thakur-Weigold from Hewlett-Packard (HP) first present some recent trends that challenge companies in the area of supply chain management and then discuss how academic research might respond to these challenges. Drawing upon HPs successful collaboration with academic institutions in the area of supply chain management, they outline a number of factors for effective interaction between industry and academia. Chapter 13, *Supply Chain Management: Interlinking Multiple Research Streams,* by James Hershauer, Kenneth Walsh, and Iris Tommelein, provides a view of the evolution of the supply chain literature that emphasizes the influence of industry, and also takes a broad view beyond a traditional operations focus.

Chapter 14, *PROFIT: Decision Technology for Supply Chain Management at IBM Microelectronics Division,* by Ken Fordyce and Gary Sullivan, provides a case history of the ongoing evolution of a major supply chain management effort in support of IBM's Technology Group. They also characterize the scope and scale of such an application, identify potential opportunities for improvement and set these within a logical evolutionary pattern, and identify research opportunities to develop new decision support capabilities. Staying with the theme of actual case studies, Young Lee and Jack Chen, in Chapter 15, *Case Studies: Supply Chain Optimization Models in a Chemical Company,* give an overview of the supply chain models that have recently been used in a large international chemical company. They describe three supply chain optimization models in detail, and discuss the lessons learned from these studies regarding issues that are especially relevant to the chemical industry.

As the foregoing descriptions indicate, the chapters in this book address a broad range of supply chain management and electronic commerce issues. The common underlying theme throughout involves the application of research to real industry contexts. The chapters are self-contained and all chapters in this book went through a thorough review process by anonymous referees. We would like to thank the chapter authors for their contributions, along with the referees, for their help in providing valuable suggestions for improvement. We would also like to thank the National Science Foundation for supporting the workshop that provided the impetus for this work (NSF Grant #DMI-0131527).

JOSEPH GEUNES, ELIF AKÇALI, PANOS PARDALOS, EDWIN ROMEIJN, AND MAX SHEN

I

SUPPLY CHAIN OPERATIONS

This page intentionally left blank

Chapter 1

COORDINATION OF INVENTORY AND SHIPMENT CONSOLIDATION DECISIONS: A REVIEW OF PREMISES, MODELS, AND JUSTIFICATION

Sila Çetinkaya
Industrial Engineering Department
Texas A&M University
College Station, Texas 77843-3131
sila@tamu.edu

Abstract This chapter takes into account the latest industrial trends in integrated logistical management and focuses on recent supply-chain initiatives that enable the integration of inventory and transportation decisions. The specific initiatives of interest include Vendor Managed Inventory (VMI), Third Party Warehousing/Distribution (3PW/D), and Time Definite Delivery (TDD) applications. Under these initiatives, substantial savings can be realized by carefully incorporating an *outbound* shipment strategy with inventory replenishment decisions. The impact is particularly tangible when the shipment strategy calls for a *consolidation* program where several smaller deliveries are dispatched as a single combined load, thereby realizing the scale economies inherent in transportation. Recognizing a need for analytical research in the field, this chapter concentrates on two central areas in shipment consolidation: i) analysis of pure consolidation policies where a shipment consolidation program is implemented on its own without coordination, and ii) analysis of integrated policies where outbound consolidation and inventory control decisions are coordinated under recent supply-chain initiatives. The chapter presents a research agenda, as well as a review of the related literature, in these two areas. Some of the recent findings of the methodological research are summarized, and current and future research endeavors are discussed. By offering a theoretical framework for modeling recent supply-chain initiatives, the chapter highlights some of the many challenging practical problems in this emerging field.

1. Introduction

1.1 Background and Terminology

This chapter concentrates on the cost saving opportunities available in *outbound* transportation. These savings are easily realizable when outbound dispatch decisions include a strategy for *shipment consolidation,* the policy under which several small loads are dispatched as a single, larger, and more economical load on the same vehicle (Brennan (1981); Hall (1987); Higginson and Bookbinder (1995)). Development of a shipment consolidation program requires strategic and operational decision-making that involves the location of consolidation terminals, development of feasible delivery routes, vehicle allocations, etc. Once higher level decisions are made, the next step is to choose an operating routine, e.g., a consolidation policy for day-to-day problems. The focus of the chapter is on analytical models for such operational decisions.

Shipment consolidation may be implemented on its own without coordination. Such a practice is called a *pure* consolidation policy. Alternatively, in choosing an operating routine, it may be useful to consider the impact of shipment consolidation on other operational decisions, such as inventory decisions. Hence, another approach is to coordinate/integrate shipment consolidation with inventory decisions. This practice is called an *integrated* inventory/shipment consolidation policy. Research on pure consolidation policies provides a foundation for the analysis of integrated models. This chapter presents a review of both of these practices, and it introduces some future research avenues in the area.

i) Pure Consolidation Policies The "operating routine" for a pure consolidation policy specifies a selected dispatching rule to be employed each time an order is received (Abdelwahab and Sargious (1990)). The relevant criteria for selecting an operating routine include customer satisfaction as well as cost minimization. Some operational issues in managing pure consolidation systems are similar to those encountered in inventory control. Two fundamental questions that must be answered are i) when to dispatch a vehicle so that service requirements are met, and ii) how large the dispatch quantity should be so that scale economies are realized. It is worth noting that these two questions relate to consolidation across time since a consolidated load accumulates by holding shipments over one or more periods. This practice is also known as temporal consolidation.

The literature on pure consolidation policies is abundant. Recent research in the area concentrates on the development of analytical models as an aid to obtaining "suitable" operating routines for temporal consolidation practices (Bookbinder and Higginson (2002); Çetinkaya and Bookbinder (2002)). However, several challenging stochastic problems remain unresolved. There is a need for additional research on identifying the structural properties of optimal pure consolidation routines and analyzing the impact of these routines on total system cost and on the timely delivery requirements of the customers.

ii) Integrated Inventory/Shipment Consolidation Policies Interest in supply-chain management arises from the recognition that an integrated plan for the chain as a whole requires coordinated decisions between different functional specialties (e.g., procurement, manufacturing, marketing, distribution). In recent years, increased emphasis has been placed on coordination issues in supply-chain research (Arntzen, Brown, Harrison, and Trafton (1995); Blumenfeld, Burns, Daganzo, Frick, and Hall (1987); Boyaci and Gallego (2002); Davis (1993); Lee and Billington (1992); Lee, Padmanabban, and Whang (1997); Stevens (1989); Tayur, Ganeshan, and Magazine (1999)). In keeping with this trend, this chapter discusses a new class of coordination problems applicable in a variety of supply-chain initiatives relying on the integration of inventory and outbound transportation decisions. Examples of these initiatives include Vendor Managed Inventory (VMI), Third Party Warehousing/Distribution (3PW/D), and Time Definite Delivery (TDD) agreements.

Revolutionized by Wal-Mart, VMI is an important coordination initiative in supply-chain management (Aviv and Federgruen (1998); Bourland, Powell, and Pyke (1996); Çetinkaya, Tekin, and Lee (2000); Kleywegt, Nori, and Savelsberg (1998); Schenck and McInerney (1998); Stalk, Evans, and Shulman (1992)). In VMI, the supplier is empowered to manage inventories of agreed-upon items at retailer locations. As a result, VMI offers ample opportunity for synchronizing outbound transportation (in particular, shipment consolidation) and inventory decisions. Similarly, 3PW/D and TDD agreements are contract based arrangements engaged in for the purpose of load optimization as well as timely delivery. The main goal of these initiatives is to design an effective distribution system.

Realization of the opportunities offered by VMI, 3PW/D, and TDD agreements, however, requires balancing the tradeoff between timely delivery and economizing on dispatch size and inventory holding costs. The integrated models discussed herein investigate these tradeoffs, and, hence, they are useful for justifying and analyzing the impact of VMI, 3PW/D, and TDD arrangements. This research has been identified through a partnership with computer and semiconductor industry members in Texas. It concentrates on identifying the properties of integrated policies and analyzing the impact of integration on cost and delivery requirements (Çetinkaya and Lee (2000); Çetinkaya and Lee (2002); Çetinkaya, Mutlu, and Lee (2002); Çetinkaya, Tekin, and Lee (2000)).

1.2 Overview

The remainder of this chapter is organized as follows. Sections 2 and 3 explain the premises and challenges of coordinating inventory and shipment consolidation decisions. While the majority of the chapter focuses on stochastic models, Section 4 provides a review of previous literature on both deterministic and stochastic models and relates it to current research endeavors in the area. Section 5 illustrates the models and methodology for some specific problems of interest. In particular, Section 5.1 concentrates on pure consolidation policies whereas Section 5.2 discusses integrated policies. The development and analysis in these sections rely on renewal theory. However, more general problems requiring the implementation of other methodologies, such as dynamic programming and stochastic programming, are also mentioned. Section 5.2 provides an introduction to the integrated models. Again, although the focus is on stochastic models, Section 5.3 presents an integrated model for the case of deterministic stationary demand. Section 5.4 focuses on integrated stochastic policies of practical interest and emphasizes the need for research on computing exact optimal policies and other extensions. Finally, Section 6 concludes the chapter.

2. Premises and Motivation

In the last few years, several competitive firms have focused on effective supply-chain practices via the new initiatives of interest in this chapter. Applied Materials, Hewlett-Packard, Compaq, and General Motors are a few examples, along with the pioneers of successful VMI practice, Wal-Mart and Procter and Gamble. As a result, the theory of coordinated inventory and transportation decisions has enjoyed a renewed interest in practical applications and academia (Bramel and Simchi-Levi (1997)). Nevertheless, most of the existing literature in the area is methodologically oriented (e.g., large scale mixed integer programs). This literature is of great value for decision making and cost optimization in a deterministic setting. However, by nature, it does not render general managerial insights into operational decisions under conditions of uncertainty or related system design issues. The research problems summarized here place an emphasis on providing insightful tools for operational decision-making and distribution system design under uncertainty. Although these problems have gained academic attention recently, there is still a need for research to meet the following objectives:

- To develop a modeling framework and theoretical understanding of inventory and transportation decisions in the context of new initiatives in supply-chain management.
- To identify optimal pure and integrated policies for general demand processes and cost structures and to develop computational procedures that simplify practical implementation.
- To analyze the cost and timely delivery implications of pure and integrated policies.
- To provide analytical tools for a comparison of different practices such as an immediate delivery policy, a pure consolidation policy, and an integrated policy.
- To render insights into effective distribution system/policy design and operational level decision-making.

The broader objective here is to explore the interaction between inventory and transportation decisions and address the question of under what conditions integration works.

Concern over the interaction between inventory and transportation costs has long been discussed in the JIT literature (Arcelus and Rowcroft (1991); Arcelus and Rowcroft (1993); Gupta and Bagchi (1987)). For illustrative purposes, let us revisit an example from Çetinkaya and Bookbinder (2002). Consider the case in Figure 1.1 where a number of small shipments arriving at origin *A* are to be delivered to destination *B*. These shipments may consist of components, or sub-assemblies, collected from various suppliers; for example, *B* might be a car assembly plant and *A* a warehouse that enables the staging JIT deliveries to *B*.

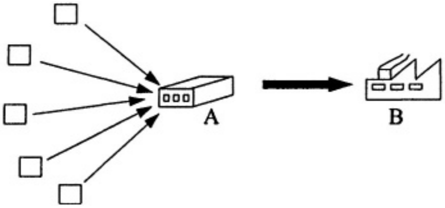


Figure 1.1. Consolidation in JIT deliveries.

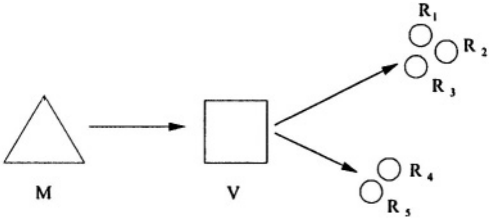


Figure 1.2. Consolidation in distribution.

On the other hand, in the context of VMI, shipment consolidation is a new area. The benefits of VMI are well recognized by successful retail businesses such as Wal-Mart. In VMI, distortion of demand information (known as the bullwhip effect) transferred from the downstream supply-chain member (e.g., retailer) to the upstream member (e.g., vendor) is minimized, stockout situations are less frequent, and system-wide inventory carrying costs are reduced. Furthermore, a VMI vendor has the liberty of controlling the downstream re-supply decisions rather than filling orders as they are placed. Thus, the approach offers a framework for coordinating inventory and *outbound* transportation decisions. The goal here is to present a class of coordination problems within this framework.

In a VMI partnership, inventory and demand information at the retailer are accessible to the vendor by using advanced on-line messaging and data retrieval systems (Cottrill (1997); Parker (1996)). By reviewing the retailers' inventory levels, the vendor makes decisions regarding the quantity and timing of re-supply. Application of VMI calls for integrating supply and outbound transportation operations through information sharing. Hence, the approach is gaining more attention as Electronic Data Interchange (EDI) technology improves and the cost of information sharing decreases.

As an example, consider the case illustrated in Figure 1.2 where *M* is a manufacturer; *V* is a vendor/distributor; and $R_i, i = 1, 2, \ldots$ is a retailer or customer. Suppose that a group of retailers (R_1, R_2, etc.) located in a given geographical region has random demands, and these can be consolidated in a larger load before a delivery is made to the region. That is, demands are not satisfied immediately, but, rather, are shipped in batches of consolidated loads. As a result, the actual inventory requirements at *V* are specified by the dispatching policy in use, and consolidation and inventory decisions at *V* should not be made in isolation from each other. In this example, the total cost for the vendor includes procurement and inventory carrying costs at *V*, the cost of waiting associated with ordered-but-not-yet-delivered demand items to the retailers, and the outbound transportation cost for shipments from *V* to the region. Also, note that while *V* is not the final destination in the supply-chain, it may be logical for various orders to be shipped together from *M* to *V,* since they will be delivered closely in time. This would be the situation if an *inbound* consolidation policy was in place. The focus of the integrated models here, however, is on outbound consolidation.

3. Modeling Challenges

Although the determination of practical decision rules for shipment consolidation has received attention in the literature, the computation of optimal policies for shipment release timing still remains an area requiring further research. In the existing literature, there are only a few guidelines for computing optimal consolidation policy parameters (Bookbinder and Higginson (2002); Çetinkaya and Bookbinder (2002); Higginson and Bookbinder (1995)). This is a challenging problem for the following reasons.

Customer Service The first complicating factor pertains to *customer service* (Çetinkaya and Bookbinder (2002)). If a temporal consolidation program is in place, then the first order received at *V* (see Figure 1.2) is from that customer who ends up waiting the longest for the goods. Thus, acceptable customer service should be assured by imposing a maximum holding time (i.e., a time-window) for the first (or any) order. Unfortunately, even after the delays of early orders are accounted for, we cannot guarantee that the subsequent order arrivals (a stochastic process) will be sufficient to achieve the low total cost sought by the consolidation strategy. Hence, research in the area should analyze cost versus the delivery time implications of different customer service levels.

Inventory Holding and Waiting Costs The second complicating factor pertains to *holding costs* and *waiting costs* (Çetinkaya and Bookbinder (2002)). Under any shipment consolidation program, some period of time elapses between the staging of a number of orders and the departure of a consolidated load. That is, shipment consolidation is implemented at the expense of customer waiting costs as well as inventory carrying costs. Holding costs represent the actual warehousing expenses during a shipment-consolidation-cycle as well as the opportunity cost in advanced payment for materials or investment in inventory. Waiting costs represent an opportunity loss in delayed receipt of revenue as well as a loss in the form of a goodwill penalty. The optimal policy, thus, should minimize the sum of the transportation, holding, and waiting costs and address the issue of balancing the three.

Interdependence of Inventory and Shipment Release Decisions If an outbound consolidation policy is in place, then *the actual inventory requirements at the vendor are partly dictated by the shipment release schedules.* Hence, coordination of inventory replenishment decisions and shipment release timings may help to reduce system-wide costs. In fact, if consolidation efforts are ignored in the optimization of inventory, the cost saving opportunities that might be realized through coordination may be overlooked. This issue is important in the context of VMI, 3PW/D, and TDD agreements where inventory decisions at the vendor account for consolidated shipments to downstream supply-chain members. Naturally, however, integrating production/inventory and shipment release timing decisions increases the problem size and complexity.

Structure of Transportation Costs Another complicating factor relates to *the structure of the transportation costs* which depend on several factors such as transportation mode, routing policies, and carriage-type. Concentrating on the case of highway transportation and ignoring the routing-related costs, let us consider the major shipping cost patterns that arise in consolidation (Hall and Racer (1995); Higginson and Bookbinder (1995); Çetinkaya and Bookbinder (2002)).

- For the case of a private-carriage, the shipping cost is primarily a function of the distance between the origin and the destination; thus, it is a fixed cost per cargo/truck for each origin-destination pair.

- When a common-carriage is used, the total shipment cost is based on the shipment quantity (total cwt.) In this case, a prototype tariff function has the form

$$c(u) = \begin{cases} d_1 u, & 0 \leq u < U_1, \\ d_{j-1} u, & U_{j-2} \leq u < U_{j-1}, j = 2, .., J \\ d_J u, & U_{J-1} \leq u, \end{cases}$$

  where $d_1 > \ldots > d_J$ denote per unit-weight freight rates, and $0 < U_1 < \ldots < U_{J-1}$ denote the break-points for shipping larger quantities.

The cost structure given by $c(u)$ implies that if $d_2 U_1 / d_1 \leq u < U_1$, then $c(u) \geq c(z)$ for all z such that $U_1 \leq z < d_1 U_1 / d_2$. However, it is unreasonable to pay more for transporting a smaller weight than a larger weight. To avoid this situation, shippers are legally allowed to over-declare the actual shipment weight. That is, the shipper has the opportunity to decrease total common-carrier charges by artificially inflating the actual shipping weight to the closest break-point (Carter, Ferrin, and Carter (1995); Higginson and Bookbinder (1995); Russell and Krajewski (1991)). The strategy of declaring "a phantom weight" is known as a *bumping clause.* Under this strategy, observe that, for example, if there is a single price-break at U_1, the *effective* common-carriage tariff function, denoted $\tilde{c}(u)$, can be represented by

$$\tilde{c}(u) = \begin{cases} d_1 u, & 0 \leq u < \hat{U}_1, \\ d_2 U_1, & \hat{U}_1 \leq u < U_1, \\ d_2 u, & U_1 \leq u. \end{cases}$$

where $\hat{U}_1 = d_2 U_1 / d_1$.

Incorporation of the bumping clause in optimization models may lead to a non-differentiable cost function. Hence, common-carrier transportation problems may be more demanding in terms of their computational requirements. With a few exceptions (Çetinkaya and Bookbinder (2002); Higginson and Bookbinder (1995); Russell and Krajewski (1991)), the concept of the bumping clause seems to be overlooked in most analytical models.

Cargo Capacity The fifth complicating factor is the effect of cargo capacity constraints. Typically, the volume of a consolidated load exceeds the cargo volume limit before an economical dispatch weight accumulates. Incorporation of cargo capacity in optimization models also leads to a non-differentiable total cost function, since, typically, cargo costs include fixed costs. Also, for stochastic problems, the weight or volume (or both) of a load accumulated during a fixed time interval is a random variable. In order to guarantee that this random variable does not exceed the existing cargo weight or volume limit, the capacity restrictions should be modeled as chance constraints, i.e., inequality constraints in the form of probabilities.

Multiple Market Areas and Products The last complication arises in coordinating shipment schedules to different market areas. The problem is particularly challenging when the demand and cost profiles for different market areas (as well as for individual customers within a given area) are different. A similar complication arises when there are multiple products. The focus in this chapter, however, is on single item, single market area problems.

It is worth noting that the above listed complications arise both in the context of pure consolidation policies and integrated policies.

4. Literature Review

4.1 Practice Oriented Literature and Applications

The principles of pure consolidation policies have traditionally been discussed in the logistics trade journals (Newbourne and Barrett (1972); Pollock (1978)). Different opportunities to realize shipment consolidation savings have also been described in fundamental logistics textbooks (Ballou (1999); Bowersox (1978); Daganzo (1996)). On the other hand, the economic justification of pure consolidation practices has received attention only in the last two decades (Blumenfeld, Burns, Diltz and Daganzo (1985); Burns, Hall, Blumenfeld and Daganzo (1985); Campbell (1990); Daganzo (1988); Gallego and Simchi-Levi (1990); Hall (1987); Higginson (1995); Pooley and Stenger (1992); Russell and Krajewski (1991); Sheffi, Eskandari, and Koutsopoulos (1988)). The early academic treatments are based on simulation models (Closs and Cook (1987); Cooper (1984); Jackson (1981); Masters (1980)). Several reasonable and easy-to-implement consolidation strategies can be identified in the previous literature. These include:

- time-based dispatch/consolidation policies, and
- quantity-based dispatch/consolidation policies.

A *time-based policy* ships accumulated loads (clears all outstanding orders) every T_c periods whereas a *quantity-based policy* ships an accumulated load when an economical dispatch quantity, say Q_c, is available. The literature also identifies a hybrid consolidation routine, called a *hybrid policy,* which is characterized by a dispatch frequency T_c and an economical dispatch quantity Q_c. Under a hybrid policy, a dispatch decision is made at $\min\{T(Q_c), T_c\}$ where $T(Q_c)$ denotes the arrival time of the Q_c^{th} demand.

Both time-based and quantity-based consolidation policies are popular in VMI, 3PW/D, and TDD applications where the interaction between inventory and shipment consolidation is considered for the purpose of cost and load optimization. Typically, time-based policies are used for A-class (lower volume/higher value) items such as commercial CPUs in the computer industry, and quantity-based policies are used for B-class and C-class (higher volume/lower value) items such as peripherals. Based on our experience, it seems that hybrid policies are not generally implemented explicitly; rather, they appear to be implicit, i.e, in managing day-to-day operations and in the troubleshooting associated with expedited orders.

Time-based and quantity-based policies are incorporated in supply contracts for the purposes of achieving timely delivery and load optimization, respectively. These contracts specify the rate schedules for TDD and full-truck-load (FTL) shipments which are also known as load-optimized deliveries. Naturally, time-based policies are suitable for TDD, whereas quantity-based policies are suitable for FTL shipments. For example, in a representative VMI application, the vendor provides warehousing and outbound transportation for finished goods and guarantees TDD and FTL shipments for outbound deliveries to the customers (i.e., a downstream supply-chain member). In this setting, since the actual inventory requirements at the vendor are dictated by the outbound shipment schedules, the inventory and outbound consolidation policies should be coordinated/integrated. We revisit this issue later in Section 5.2 where we also discuss a related modeling methodology.

4.2 Quantitative Literature

4.2.1 Simulation and Analytical Models for Pure Consolidation Policies. Higginson and Bookbinder (1994) compare time-based, quantity-based, and hybrid policies in a pure consolidation setting via a simulation model where most of the relevant parameters are varied. However, the *optimal* choices for Q_c and T_c may not be among the values tested for any of the policies. Although this limitation of simulation has been recognized in the early literature, there are only a few papers that provide analytical models for shipment release timing. Higginson and Bookbinder (1995) employ a Markovian Decision Process model to compute the optimal quantity policies *numerically.* Gupta and Bagchi (1987) adopt Stidham's (Stidham (1977)) results on stochastic clearing systems which are characterized by stochastic input (e.g., freight from M to V in Figure 1.2) and an output mechanism (e.g., dispatching a vehicle from V to the final destination in Figure 1.2) that clears the system (e.g., V in Figure 1.2). Brennan (1981) obtains structural results when consolidated loads are reviewed on a periodic basis for both deterministic and stochastic demand problems. Other analytical treatments of pure consolidation policies include those based on queueing theory in the setting of passenger transport and dynamic vehicle dispatch (Gans and van Ryzin (1999); Minkoff (1993); Powell (1985); Powell and Humblet (1986)). One common characteristic of the previous studies is that they focus mainly on quantity policies and do not consider compound demand processes. In a recent paper, Çetinkaya and Bookbinder (2002) model compound input processes and analyze both private-carriage and common-carriage problems. We revisit this work later in Section 5.1 where we also provide a list of future research issues in pure consolidation policies. Nevertheless, all of the papers mentioned so far in this section concentrate on pure consolidation policies, ignoring the following:

- the interaction between inventory and shipment consolidation decisions,
- cargo capacity constraints, and
- multiple market area distribution problems.

4.2.2 Analytical Models for Integrated Policies. Although the literature on integrated inventory and transportation decisions is abundant, most of the existing work is methodologically oriented and concentrates on algorithmic procedures for large scale optimization models. Furthermore, with a few exceptions (e.g., Çetinkaya and Lee (2000)), the existing literature does not directly address the effects of temporal consolidation. Bramel and Simchi-Levi (1997) provide an excellent review of the literature on integrated models for inventory control and vehicle routing (also see, Anily and Federgruen (1990); Anily and Federgruen (1993); Chan, Muriel, Shen, Simchi-Levi, and Teo (2002); Federgruen and Simchi-Levi (1995); Hall (1991); Viswanathan and Mathur (1997)).

In general the multi-echelon inventory literature and, in particular, the problem of buyer-vendor coordination is closely related to the integrated problems considered here. For example, Axsäter (2000); Banerjee (1986); Banerjee (1986); Banerjee and Burton (1994); Goyal (1976); Goyal (1987); Goyal and Gupta (1989); Joglekar (1988); Joglekar and Tharthare (1990); Lee and Rosenblatt (1986) and Schwarz (1973) present meritorious results in this area. However, the previous work in buyer-vendor coordination neglects the complicating factors of shipment consolidation addressed in Section 2 and throughout this chapter.

Inventory lot-sizing models in which transportation costs are considered explicitly are more distantly related to the topic of interest in this chapter. In recent years, joint quantity and freight discount problems have received significant attention in logistics research (Aucamp (1982); Baumol and Vinod (1970); Carter and Ferrin (1996); Constable and Whybark, (1978); Diaby and Martel (1993); Gupta (1992); Hahm and Yano (1992); Hahm and Yano (1995a); Hahm and Yano (1995b); Henig, Gerchak, Ernst, and Pyke (1997); Knowles and Pantumsinchai (1988); Lee (1986); Lee (1989); Popken (1994); Sethi (1984); Tersine and Barman (1991); Tyworth (1992)). The efforts in this field are mainly directed towards deterministic modeling with an emphasis on *inbound* transportation. These previous models do not address the outbound distribution issues that arise, particularly in the context of VMI, 3PW/D, and TDD arrangements.

4.2.3 Limitations. Although a large body of literature in the general area of integrated inventory and transportation decisions exists, the following research problems need attention:

- Computation of parameter values for *practical* pure consolidation policies (e.g., time, quantity, and hybrid policies) in a stochastic setting for general demand processes, for transportation by a private fleet and common-carriage trucking company under cargo capacity constraints, and for single and multiple market area problems.

- Characterization of *optimal* pure consolidation policies for general demand processes, for transportation by a private fleet and common-carriage trucking company under cargo capacity constraints, and for single and multiple market area problems.

- Computation of parameter values for *practical* integrated policies in a stochastic setting for general demand processes, for transportation by a private fleet and common-carriage trucking company under cargo capacity constraints, and for single and multiple market area problems.

- Characterization of *optimal* integrated policies and integrated policies which assure acceptable customer service, again, in a stochastic setting for general demand processes, for transportation by a private fleet and common-carriage trucking company under cargo capacity constraints, and for single and multiple market area problems.

- Analysis of time versus cost tradeoffs for pure and integrated policies and analysis of conditions under which integration works best.

- Development of analytical as well as sophisticated simulation tools for comparison of different practices and initiatives that render insights into distribution system/policy design and operational level decision making.

5. Models and Methodology

Some special cases of the problems itemized above have been modeled and solved in Çetinkaya and Bookbinder (2002); Çetinkaya and Lee

(2000); Çetinkaya and Lee (2002); Çetinkaya, Mutlu, and Lee (2002), and Çetinkaya, Tekin, and Lee (2000), and an overview of these results is presented next. Several important future research directions are also discussed.

5.1 Pure Consolidation Policies

5.1.1 Problem Setting. To set the stage for a mathematical formulation, we revisit the example illustrated in Figure 1.1 where a number of small shipments arriving at *A* are to be delivered to *B*. Consider the case where the arrival times, as well as the weights of the shipments, are random variables. The purpose is to find a consolidation policy that minimizes the expected long-run average cost of shipping plus holding shipments at *A*. The consolidation policy parameters specify i) when to dispatch a vehicle from *A* so that service requirements are met, and/or ii) how large the dispatch quantity should be so that scale economies are realized.

A similar pure consolidation problem is also encountered at *V* in Figure 1.2. Suppose that a group of customers (e.g., R_1, R_2, and R_3) located in a given market area places small orders of random size at random times, and suppose these can be consolidated in a larger load before a delivery truck is sent to the market area. Note that in this latter example, the objective function should include the cost of waiting associated with ordered, but not-yet-delivered, demand items.

If dispatch decisions (at *A* or *V*) are made on a recurring basis, under certain additional assumptions, we may utilize renewal theory. More specifically, provided that the underlying order arrival and dispatch processes satisfy certain conditions, we may compute the parameter values for time, quantity, and hybrid policies through renewal theoretic analysis. A simple example of such an analysis, based on the results in Çetinkaya and Bookbinder (2002), is presented in the following discussion.

Recall that a time-based dispatch policy ships each order (consolidated or not) by a predetermined shipping date. In turn, a stationary time-based policy is characterized by a single parameter, say T_c, which is called the *critical (maximum) holding time.* A second approach is to employ a quantity-based policy under which a dispatch decision is made when the accumulated load is more than a *minimum (critical) weight,* say Q_c. Finally, the third approach is a combination of the above two. A hybrid policy aims to dispatch all orders before a predetermined shipping date (during a time-window); but if a minimum consolidated load accumulates before that date, then all outstanding orders are dispatched immediately. On the other hand, if a minimum consolidated weight does not accumulate in time, all orders are dispatched on the prespecified date. Economies of scale associated with shipping a larger quantity may be sacrificed; however, customer service requirements are always met. Concentrating on single market area problems, Çetinkaya and Bookbinder (2002) report results on computing the parameters of optimal time and optimal quantity policies separately for the cases of *private-carriage* and *common-carriage.* These existing results make some specific assumptions about the underlying demand processes as we explain shortly.

5.1.2 An Illustrative Model for Private-Carriage.

Considering a single market area problem, an illustration for computing the critical holding time T_c under a *pure* time-based policy for the consolidation problem at *V* (Figure 1.2) is presented below. That is, we ignore the inventory replenishment and carrying costs at *V* and concentrate only on the outbound dispatch problem. Later, in Section 5.2, we discuss the integrated policy where inventory at *V* is modeled explicitly.

Suppose that orders from customers located in a given market area form a stochastic process with interarrival times $\{X_k : k = 1, 2, \ldots\}$. For the moment, assume that $X_k \geq 0, k = 1, 2, \ldots$ are independent and identically distributed (i.i.d.) according to distribution function $F(\cdot)$, where $F(0) < 1$, and density $f(\cdot)$. Letting $S_0 = 0$ and $S_k = \sum_{i=1}^{k} X_i$, we define $N_1(t) = \sup\{k : S_k \leq t\}$. It follows that $N_1(t)$ is a renewal process that registers the number of orders placed by time t. Also, let $\{W_k : k = 1, 2, \ldots\}$ represent another sequence of i.i.d. random variables with density $g(\cdot)$ and distribution $G(\cdot)$ where $G(0) < 1$. We shall interpret W_k as the weight of the k^{th} order. Thus, $\mathcal{N}(t) = \sum_{i=1}^{N_1(t)} W_i$ is the weight of the cumulative demand until time t. We define $D_0 = 0$, $D_k = \sum_{i=1}^{k} W_i$, and $N_2(q) = \sup\{k : D_k \leq q\}$. Likewise, $N_2(q)$ is a renewal process and registers the number of orders until the cumulative demand reaches q. Since a dispatch decision is taken every T_c days, the maximum holding time, T_c, correspondingly represents the length of a shipment-consolidation-cycle for the market area. A realization of the demand process under this scenario is depicted in Figure 1.3.

Let $L(t)$ represent the size of the consolidated load, i.e., the number of outstanding demands, at time epoch t. The consolidation system is cleared and a new shipment-consolidation-cycle begins every T_c time-units. In turn, $L(jT_c), j = 1, 2, \ldots$ is a sequence of random variables representing the dispatch quantities. Keeping this observation in mind, we define

$$\mathcal{N}_j(T_c) \equiv L(jT_c), \quad j = 1, 2, \ldots.$$

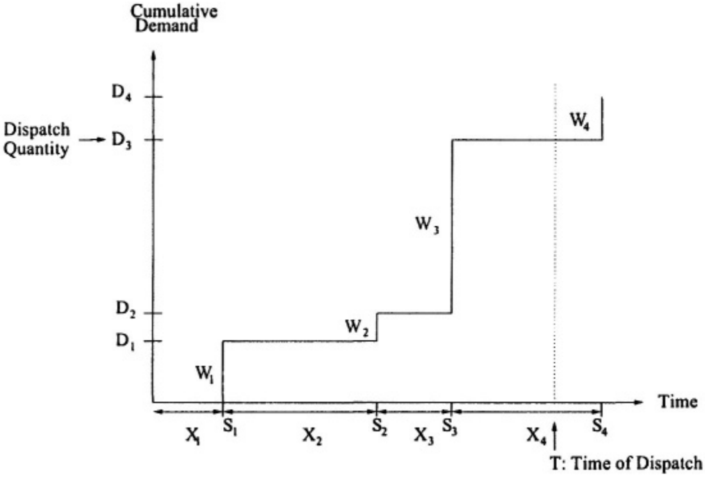


Figure 1.3. A realization under a time-based policy.

Observe that the sequence $\mathcal{N}_j\,(T_c)\,, j\;=\;1,2,\ldots$ symbolizes the dispatch/shipment release weights under the time-based dispatching policy in use whereas the sequence $W_n, n = 1, 2, \ldots$ represents the actual order weights. The process $\mathcal{N}_j\,(T_c)\,, j = 1, 2, \ldots$ is a function of T_c, and thus the dispatch quantities are random variables established by the parameter of the shipment-consolidation policy in use.

If $\mathcal{N}(t)$ is a compound Poisson process, then the random variables $\mathcal{N}_j\,(T_c)\,, j = 1, 2, \ldots$ are i.i.d., each having the same distribution as the random variable $\mathcal{N}\,(T_c)$. It is worth noting that for other renewal processes, $\mathcal{N}_j\,(T_c)\,, j = 1, 2, \ldots$ are not necessarily i.i.d., and this is a major source of difficulty for the problem of interest. Obtaining analytical results for general renewal processes seems to be rather challenging if not impossible, and it remains an open area for future investigation. Here, as in Çetinkaya and Bookbinder (2002), the focus is on the case of compound Poisson processes for analytical tractability.

The expected long-run average cost, denoted by $\mathcal{C}(T_c)$, is computed using the renewal reward theorem, i.e.,

$$\begin{aligned}\mathcal{C}(T_c) \;&=\; \frac{E[\text{Transportation cost per shipment-consolidation-cycle}]}{T_c}\\ &\quad+\frac{E[\text{Waiting cost per shipment-consolidation-cycle}]}{T_c}.\end{aligned}$$

If truck capacity constraints are ignored and only a fixed transportation cost, denoted by K_c, is associated with a dispatch decision, then

$$C(T_c) = \frac{K_c + E[\text{Waiting cost per shipment-consolidation-cycle}]}{T_c}.$$

Figure 1.4 illustrates the accumulation of waiting costs in an arbitrary consolidation cycle. For the particular consolidation-cycle illustrated in Figure 1.3, $N_1\,(T_c) = 3$ so that the corresponding waiting cost is given by

$$w\left[(X_2 D_1 + X_3 D_2) + (T_c - S_3) D_3\right]$$

$$= w\left[X_2 W_1 + X_3(W_1 + W_2) + (T_c - S_3)(W_1 + W_2 + W_3)\right],$$

where w denotes the cost of waiting for one unit of demand per unit-time. Using these illustrations and letting $A_1\,(T_c) = T_c - S_{N_1(T_c)}$ denote the age of $N_1(t)$ at T_c, it can be easily verified that

$$E[\text{Waiting cost per shipment-consolidation-cycle}] =$$

$$wE\left[\sum_{i=2}^{N_1(T_c)} X_i \sum_{j=1}^{i-1} W_j\right] + wE\left[A_1\,(T_c)\,\mathcal{N}\,(T_c)\right].$$

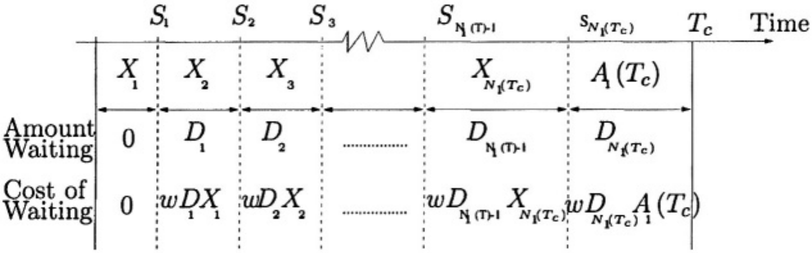


Figure 1.4. Amount waiting under a time-based policy.

Also,

$$E\left[\sum_{i=2}^{N_1(T_c)} X_i \sum_{j=1}^{i-1} W_j\right] = E[W_k]E\left[v(T, N_1\,(T_c))\right], \quad \text{and}$$

$$E\left[A_1\,(T_c)\,\mathcal{N}\,(T_c)\right] = Z\,(T_c),$$

where

$$
\begin{aligned}
v(t,k) &= \sum_{i=1}^{k-1} iE[X_{i+1}|N_1(t)=k], \\
Z(t) &= z(t) + \int_0^t Z(t-y)dF(y), \quad \text{and} \\
z(t) &= E[W_k]\int_0^t E[A_1(t-y)]dF(y).
\end{aligned}
$$

When demand is a compound Poisson process, then it is easy to obtain an analytical expression for T_c. The result, a variation of the EOQ formula, is not surprising. As we have mentioned earlier, for other compound processes, the optimal T_c cannot be computed using renewal theory; more general approaches such as Markov renewal theory, Markov decision processes, or stochastic dynamic programming are feasible. However, these approaches have not yet been investigated. The above approach can also be applied to obtain an optimal quantity-based policy for which similar results are applicable not only for compound Poisson processes but also for compound renewal processes. For those computations, $N_2(q)$ plays the role of $N_1(t)$ in computing a time-based policy. The following comparative results for compound Poisson processes are based on analysis of the private-carriage case in Çetinkaya and Bookbinder (2002).

PROPERTY 1.1

i) The expected dispatch quantity under the optimal time-based policy is larger than the optimal critical weight but smaller than the mean load dispatched under the optimal quantity-based policy.

ii) An optimal quantity-based policy has a mean shipment-consolidation-cycle length larger than that of the corresponding optimal time-based policy. Hence, the time-based policy offers superior service to customers, not only in the sense that a specific delivery time can be quoted, but also in the sense that delivery frequencies are higher.

5.1.3 An Illustrative Model for Common-Carriage.

It is the cost structure and parameters that distinguish between the common and private carriage. However, as we illustrate next, this distinction leads to an important computational difficulty in obtaining an expression for expected transportation cost per shipment-consolidation-cycle so that insightful structural results cannot be presented even for the simpler case of compound Poisson demand processes.

Let us consider the case of a single price-break so that the effective common-carriage cost function is given by $\tilde{c}(u)$ in Section 3. For illustrative purposes, let us try to compute the optimal T_c for a common-carriage under the assumptions that X_k and W_k are exponentially distributed with respective parameters λ and α (or $1/E[X_k]$ and $1/E[W_k]$). The expected waiting cost per shipment-consolidation-cycle is computed as in the case of the private-carriage. For the specific example under consideration, $\mathcal{N}(t)$ is a compound Poisson process. Thus, it is straightforward to show that

$$\frac{E[\text{Waiting cost per shipment-consolidation-cycle}]}{T_c} = \frac{wE[W_k]\lambda T_c}{2}.$$

However, the remaining terms of $\mathcal{C}\left(T_c\right)$ require the density function of $\mathcal{N}(t)$, denoted by $\phi(t,x)$, for which a closed form expression does not exist in most cases. The expected transportation cost per shipment-consolidation-cycle is given by

$$E\left[\tilde{c}\left(\mathcal{N}\left(T_c\right)\right)\right] = \int_0^\infty E\left[\tilde{c}(x)\right]\phi\left(T_c,x\right)dx$$

$$= \int_0^{\hat{U}_1} d_1 x\phi\left(T_c,x\right)dx + \int_{\hat{U}_1}^{U_1} d_2 U_1 \phi\left(T_c,x\right)dx + \int_{U_1}^{\infty} d_2 x\phi\left(T_c,x\right)dx,$$

Based on the result in Medhi (1994) (p. 176-7), the probability generating function (p.g.f.) of $\mathcal{N}(t)$, denoted $P_\phi(\cdot)$, is given by $P_\phi(s) = \exp\left[\lambda t\left(P_g(s)-1\right)\right]$, where $P_g(s)$ denotes the p.g.f. of W_k. However, this approach does not lead to a closed form. expression for $\phi(t,x)$, except for the special case where W_k is geometric. The above expression for $P_\phi(s)$ can be used to evaluate $\phi(t,x)$ *numerically,* and then the outcome can be utilized for a numerical evaluation of $\mathcal{C}\left(T_c\right)$. This, in turn, requires the use of numerical integration procedures. Although such an approach is feasible, the corresponding computations may require some effort. Easier-to-compute approximations are presented in Çetinkaya and Bookbinder (2002). Nevertheless, even if we can numerically evaluate $\mathcal{C}\left(T_c\right)$, its optimization requires an enumeration approach. That is, there is no guarantee that the global optimum can be found in a reasonable amount of time because, depending on the model parameters, the cost function may not be convex.

Again, the above approach can also be applied to obtain an optimal quantity-based policy under a common-carriage tariff function. Although a closed form solution does not exist, the numerical computations may be easier. This is because, to obtain an optimal quantity-based policy, we do not need an expression for $\phi(t,x)$ but rather an expression for the density function of the "excess life" at q for the process $N_2(q)$. As before, there is no guarantee that the global optimum can be found in a reasonable amount of time unless the cost function is convex for the specific values of the model parameters under consideration.

5.1.4 A Simplified Common-Carriage Model for Poisson Demands. As a simpler special case, assume that demand follows a pure (unit) Poisson process with parameter λ, and let us illustrate the computation of an optimal time-based policy for the case of the single price-break . The individual orders are of size one-unit (i.e., they are no longer random variables), and hence the consolidated order weight during a cycle time of T_c, $\mathcal{N}(T_c)$, is a Poisson random variable with parameter λT_c. Then, the density and distribution of $\mathcal{N}(T_c)$, denoted by $\phi(T_c, x)$ and $\Phi(T_c, x)$, respectively, are given by

$$\phi(T_c, x) = \frac{\exp(-\lambda T_c)(\lambda T_c)^x}{x!}, \text{ and } \Phi(T_c, x) = \sum_{k=0}^{x} \frac{\exp(-\lambda T_c)(\lambda T_c)^k}{k!}.$$

Since $\mathcal{N}(T_c)$ is now a discrete random variable, the expected transportation cost per shipment-consolidation-cycle is given by

$$E[\tilde{c}(\mathcal{N}(T_c))] = \sum_{x=0}^{\hat{U}_1} d_1 x \phi(T_c, x) + \sum_{x=\hat{U}_1+1}^{U_1} d_2 U_1 \phi(T_c, x)$$
$$+ \sum_{x=U_1+1}^{\infty} d_2 x \phi(T_c, x).$$

It follows that

$$C(T_c) = \frac{h\lambda T_c}{2} + \frac{d_1}{T_c} \sum_{x=0}^{\hat{U}_1} x\phi(T_c, x) + \frac{d_2 U_1}{T_c} \sum_{x=\hat{U}_1+1}^{U_1} \phi(T_c, x)$$
$$+ \frac{d_2}{T_c} \sum_{x=U_1+1}^{\infty} x\phi(T_c, x).$$

If one can assume that T_c is large enough so that the probability of an empty dispatch is zero, i.e., $Pr\{\mathcal{N}(T_c) = 0\} \approx 0$, then the above expression for $C(T_c)$ can be further simplified leading to

$$C(T_c) = \frac{h\lambda T_c}{2} + \lambda[d_1 \Phi(T_c, \hat{U}_1 - 1) - d_2 \Phi(T_c, U_1 - 1)]$$

$$+\frac{d_1}{T_c}[\Phi(T_c, U_1) - \Phi(T_c, \hat{U}_1)] + d_1\lambda.$$

For any given T_c, the four functional values of $\Phi(\cdot)$ can be easily obtained using a Poisson distribution table. Therefore, this expression for $\mathcal{C}(T_c)$ is obviously easier to evaluate and optimize relative to the more general case. For all practical purposes, given the specific values of the model parameters for the consolidation system under consideration, a numerical solution can be computed in a straightforward fashion, e.g., through a numerical evaluation of $\mathcal{C}(T_c)$ on a spreadsheet. Since we have a closed form expression of $\mathcal{C}(T_c)$, it is easier to check the convexity or unimodality of this function for a given parameter set.

For the case of a quantity policy, if demand can be modeled as a pure Poisson process, there is no risk of overshooting the target weight Q_c. That is, the consolidated weight per cycle is no longer a random variable, and it is exactly equal to Q_c. Keeping this observation in mind, the analysis is straightforward because

$$E[\text{Transp. cost per consolidation-cycle}] = \begin{cases} d_1 Q_c, & Q_c \leq \hat{U}_1, \\ d_2 U_1, & \hat{U}_1 < Q_c \leq U_1, \\ d_2 Q_c, & Q_c > U_1 \end{cases}$$

Also, since interarrival times are exponential with parameter λ and mean $E[X_k] \equiv 1/\lambda$

$$\begin{aligned} E[\text{Waiting cost per consolidation-cycle}] &= wE\left[\sum_{i=1}^{Q_c-1} X_{i+1}\right] \\ &= \frac{wE[X_k]\,Q_c(Q_c-1)}{2}. \\ E[\text{consolidation-cycle length}] &= E[X_1 + X_2 + \ldots + X_{Q_c}] \\ &= E[X_k]\,Q_c. \end{aligned}$$

It follows that

$$\mathcal{C}(Q_c) = \frac{h(Q_c-1)}{2} + \begin{cases} \frac{d_1}{E[X_k]}, & Q_c \leq \hat{U}_1, \\ \frac{d_2 U_1}{E[X_k]Q_c}, & \hat{U}_1 < Q_c \leq U_1, \\ \frac{d_1}{E[X_k]}, & Q_c > U_1, \end{cases}$$

and hence the optimal quantity-based policy is easy to compute.

5.1.5 Extensions, Variations, Research Issues.

Among the three practical policies of interest for pure consolidation (i.e., time-based, quantity-based, and hybrid policies), computing the parameters of the optimal hybrid policy is the most complex, even under some restrictive assumptions on the demand processes. Work on the application of renewal theory to obtain an optimal hybrid policy (Çetinkaya and Bookbinder (1997)) will be reported in a subsequent publication. Computing the optimal parameters of practical pure consolidation policies for multiple market area/product problems and the finite cargo capacity problem are important future research directions. In addition to calculating the optimal policy parameters, more research is needed for studying the specific conditions under which each policy is preferable, characterizing the differences in shipment-consolidation-cycle length and mean dispatch quantity for those policies, and analyzing the respective variances.

For the purpose of incorporating other realistic problem characteristics, such as general order arrival patterns and cargo capacity constraints, into an analytical model, some approximations may be necessary. For validation purposes, the resulting approximate models should be tested via computer simulation on a set of problems representing a large set of "reasonable" operating parameters. Empirically observed performance measures (e.g., delivery times, costs, and service levels) can then be compared with the models' estimates. Such simulation models are aimed at generating point and interval estimates for the performance measures of interest in a real consolidation system operating under the assumptions of the analytical models.

5.2 Coordinated/Integrated Models

Problem Setting Again, recall the consolidation problem at V (Figure 1.2), but suppose that in order to optimize the inventory and shipment consolidation decisions simultaneously, we wish to compute the parameters of an integrated policy. These parameters determine: i) how often to dispatch a truck so that transportation scale economies are realized and timely delivery requirements are met, and ii) how often, and in what quantities, to replenish the inventory at V.

Suppose that the vendor V replenishes its inventory for a single item from an external source M with ample supply, carries inventory, and realizes a sequence of random demands in random sizes from a group of downstream supply-chain members located in a given market area. In order to benefit from the cost saving opportunities associated with shipping a larger consolidated load, the vendor takes the liberty of not delivering small orders immediately. That is, the vendor may adopt one of the three practical shipment consolidation policies introduced earlier. Shipment consolidation may lead to substantial savings in outbound transportation costs; but, in this case, it is implemented at the expense of customer waiting costs as well as inventory carrying costs. As we have mentioned earlier, waiting costs represent an opportunity loss in delayed receipt of revenue as well as a loss in the form of a customer goodwill penalty. With higher volume/lower value items, a consolidated outbound load accumulates faster, so waiting time is not excessive. In turn, transportation scale economies may easily justify the waiting and inventory holding costs due to consolidation and inventory kept in the vendor's warehouse. Provided that the waiting time is reasonable, i.e., delivery time-window requirements are satisfied, the downstream supply-chain member (e.g., retailer) may agree to wait under the circumstances that their shelf space for carrying extra stock is limited or that carrying inventory for certain items is not desirable. Therefore, there is no inventory (and, hence, no inventory holding costs) at the downstream supply-chain member. As an example, consider a product that is clearly unreasonable for the retailer to keep in stock—say office photocopy machines or expensive laptop computers. The retailer may own some display models of products, but, typically, acts as a catalog or internet sales agent who helps customers decide what type of product best suits their needs and offers after-sales service so that customer orders are placed from and delivered to retail locations. These situations are particularly common for businesses selling high-tech, bulky, or expensive items through retail stores where inventory holding cost for the retailer is high and waiting cost is modest for "reasonable" time intervals so that the retailer does not carry inventory.

As we mentioned earlier, this class of problems arises in the context of VMI and 3PW/D programs. Under these programs, the vendor is authorized to manage inventories of agreed upon stock-keeping-units at a downstream supply-chain member, e.g., a distributor, a retail location, or a customer. By retrieving demand information at the downstream supply-chain site, the vendor makes decisions regarding the quantity and timing of re-supply. Under this arrangement, the vendor has the autonomy to hold small orders until an economical dispatch quantity (i.e., a large outbound load realizing transportation scale economies) accumulates. As a result, the actual inventory requirements at the vendor are in part specified by the parameters of the outbound shipment release policy in use.

The term *vendor* is used loosely here; depending on the industry, it may represent a manufacturer or a distributor simply taking advantage of a possible cost saving opportunity through coordinating inventory and outbound transportation decisions. For example, in the computer industry in Texas, a VMI vendor is typically a third party logistics company which is in charge of the warehousing and distribution programs of a manufacturer. That is, the third party carries inventory of finished goods (e.g., consumer and commercial CPUs) and peripherals (e.g., speakers, printers, etc.) at its warehouse and arranges outbound transportation for replenishing stock at a downstream supply-chain member who typically does not carry any inventory other than display models.

In recent years, time-based shipment consolidation policies have become a part of transportation contracts between partnering supply-chain members. These contracts, also known as TDD agreements, are common between third party logistics service providers and their partnering manufacturing companies. In a representative practical situation, a third party logistics company provides warehousing and transportation for a manufacturer and guarantees TDD for outbound deliveries to customers. Such an arrangement is particularly useful for effective VMI where the “vendor,” or its third party representative, manages the inventory replenishment and outbound transportation decisions at the “vendor’s outbound warehouse,” and customers are willing to wait under the terms of a contract at the expense of some waiting costs for the “vendor.”

For illustrative purposes, in Section 5.3 we discuss an elementary deterministic model that serves as a basis for computing an integrated policy applicable for the problem setting described above. The specific stochastic problems of interest are discussed in Section 5.4.

5.3 Deterministic Models for Coordinated/Integrated Policies

In the interest of simplicity, let us assume that the aggregate demand rate for the market area, denoted by *D,* is known and constant. Also, let T_r and Q_r denote the length of an inventory-replenishment-cycle and the replenishment order quantity for the vendor, respectively. Hence, $Q_r = DT_r$.

Consider a vendor who consolidates shipments before delivering individual small orders. The problem is to identify a reasonable integrated policy for inventory replenishment and outbound dispatch. A naïve policy of this type is characterized by two parameters, n and T_c, and it is called an (n, T_c) policy. Here, T_c denotes the length of a shipment-consolidation-cycle whereas n denotes the number of dispatch decisions within T_r. Considering the case where pre-shipping to the retailer is prohibited and inventory at the vendor is not allowed to go below zero, we have $T_r = nT_c$. Let Q_c denote the size of a consolidated load accumulated during T_c, i.e., Q_c denotes the dispatch quantity. It follows

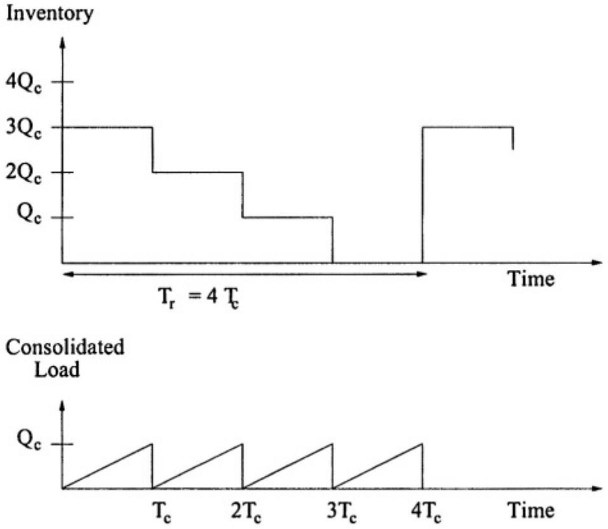


Figure 1.5. Joint consolidation and replenishment decisions: Deterministic demand problem.

that $Q_c = DT_c$. This naïve policy (under which inventory and consolidated load profiles are depicted in Figure 1.5) is a time-based dispatch policy where an outbound dispatch is made every T_c time periods. However, since demand is deterministic, the time-based dispatch policy with parameter T_c is equivalent to the quantity-based dispatch policy with parameter $Q_c = DT_c$. Unfortunately, this simplifying property does not hold for stochastic demand problems.

Suppose that K_r denotes the fixed cost of a replenishment; c_r denotes the unit procurement cost; K_c denotes the fixed cost of dispatching a truck; h denotes the inventory carrying cost per unit per unit of time; w denotes the customer waiting cost per unit per unit of time; and $C(n, T_c)$ denotes the long-run average cost per unit of time. If truck capacity constraints (i.e., cargo capacity constraints) are ignored, then

$$C(n, T_c) = c_r D + \frac{K_r}{nT_c} + \frac{K_c}{T_c} + \frac{h(n-1)DT_c}{2} + \frac{wDT_c}{2}.$$

5.3.1 An Approximate Solution. Although the above function is not jointly convex in n and T_c, by using first order optimality conditions, we can show that it has a unique minimum. We can also obtain closed form expressions for optimal n and T_c values. We define

$$n_0 = \sqrt{\frac{K_r(w-h)}{K_c h}},$$

and we let (n^*, T_c^*) denote the global minimizer of $C(n, T_c)$. If $w \leq h$, then

$$n^* = 1 \text{ and } T_c^* = \sqrt{\frac{2(K_r + K_c)}{Dw}}.$$

Otherwise, n^* is either $\lfloor n_0 \rfloor$ or $\lceil n_0 \rceil$ depending on which one yields a lower value of $C\left(n, T_c(n)\right)$ where

$$T_c(n) = \sqrt{\frac{2(K_r + nK_c)}{Dn[h(n-1)+w]}}.$$

Once n^* and T_c^* are computed, the optimal inventory-replenishment-cycle length, the replenishment quantity, and the dispatch quantity are computed using the basic relations $T_r = nT_c$, $Q_r = DT_r$, and $Q_c = DT_c$. The above result implies that if the cost of waiting is less than the cost of holding, then there is no incentive to carry inventory at the vendor's warehouse, and, hence, it is operated as a cross-docking terminal. On the other hand, if the cost of holding is less than the cost of waiting, then the warehouse is operated as a break-bulk terminal where stock is replenished in bulk, inventory is carried, and several outbound shipments are dispatched in a replenishment-cycle.

It is important to note that, for the problem of interest, n^* and T_c^* cannot guarantee an optimal policy within the class of all possible policies. This is because in computing n^* and T_c^*, we assume $T_r = nT_c$. In other words, we compute n^* and T_c^* by restricting our attention to the class of stationary shipment consolidation policies. It can be shown that (see Çetinkaya and Lee (2002)), under the exact optimal policy, the times between successive dispatch decisions are non-decreasing, i.e., not necessarily constant. Therefore, a stationary policy is not optimal in general. Furthermore, under the exact optimal policy, the question of "whether the warehouse should be operated as a cross-docking point or a break-bulk terminal" not only depends on the values of w and h, but also on the values of K_r and K_c (see Theorem 1.4 below).

REMARK 1.2 *Although the motivations of the underlying problems are substantially different, the similarity between the model discussed in this section and the multi-echelon inventory model studied in Schwarz's (Schwarz (1973)) seminal paper should be noted (also see Hahm and*

Yano (1992) that take into account transportation costs and provide a generalization of Schwarz (1973)). The mathematical formulation presented in this section is similar to Schwarz's formulation for the deterministic, one-warehouse, one-retailer problem. However, this formulation leads to an exact *optimal solution for Schwarz's problem, whereas it only gives an* approximate *solution for our problem. In fact, our purpose for discussing this simple deterministic model is to illustrate the resemblance, as well as the difference, between a simple multi-echelon inventory model and a simplified deterministic version of the general problem of interest in this chapter.*

5.3.2 Exact Optimal Solution.

The exact optimal solution of the problem satisfies the following property (Çetinkaya and Lee (2002)).

PROPERTY 1.3

i) Under the optimal policy, if there are more than two shipment-consolidation-cycles within a replenishment-cycle, then each consolidation-cycle is of equal length with the exception of the last one.

ii) Under the optimal policy, if there is more than one shipment-consolidation-cycle within a replenishment-cycle, then the last shipment-consolidation-cycle is the longest one.

It follows that each replenishment-cycle consists of n shipment-consolidation-cycles of length T_1 and one last shipment-consolidation-cycle of length T_2 where $T_1 \leq T_2$. Hence, $T_r = nT_1 + T_2$ so that each replenishment cycle consists of $n+1$ shipment-consolidation-cycles where $n = 0, 1, \ldots$. Based on these results, the average annual total cost can be expressed as a function of n, T_1, and T_r. Let $C(n, T_1, T_r)$ denote the average annual total cost in this case. It is easy to show that

$$C(n, T_1, T_r) = c_r D + \frac{wDT_r}{2} + \frac{2K_r + 2(n+1)K_c + D(n+1)nT_1^2(h+w)}{2T_r}$$
$$-wDnT_1.$$

Let n^*, T_1^*, and $T_r{}^*$ denote the optimal solution, and let T_2^* denote the corresponding optimal T_2 value. Also, let

$$\hat{n}_0 = \sqrt{\frac{wK_r}{hK_c}} - 1, \quad A(n) = D^2 w(h+w)n(n+1),$$

$$B(n) = 2Dw[K_c(n+1) + K_r], \quad \text{and} \quad C(n) = Dwn.$$

THEOREM 1.4 *If $wK_r \leq hK_c$ then*

$$n^* = 0, \quad T_1^* = 0, \quad \textit{and } T_r{}^* = \sqrt{\frac{2(K_c + K_r)}{Dw}}.$$

Otherwise,

$$n^* = \arg\min\left\{\mathcal{C}(\lfloor \hat{n}_0 \rfloor), \mathcal{C}(\lceil \hat{n}_0 \rceil)\right\}, \quad T_1^* = \sqrt{\frac{C(n^*)^2 B(n^*)}{A(n^*)^2 - C(n^*)^2 A(n^*)}},$$

$$\textit{and} \quad T_r{}^* = \frac{1}{Dw}\sqrt{\frac{A(n^*)B(n^*)}{A(n^*) - C(n^*)^2}}.$$

5.3.3 Capacitated Problem. Based on the results of the uncapacitated problem above, in Çetinkaya and Lee (2002) efficient algorithms are also developed for computing the integrated policy parameters for finite cargo capacity problems, i.e., the case where a fixed cost is associated with each truck having a finite capacity and the number of trucks dispatched at the end of a consolidation-cycle is a decision variable. These algorithms can easily be applied for solving buyer-vendor inventory problems, e.g., the one-warehouse/one-retailer case in Schwarz (1973) with cargo capacity constraints. Hence, the solution techniques developed are directly applicable for a broader class of problems. The following theorem describes important structural properties of the problem under capacity constraints.

PROPERTY 1.5 *The capacitated problem satisfies the following two properties:*

i) Under the optimal policy, if there are more than two shipment-consolidation-cycles within a replenishment-cycle, then each consolidation-cycle is of equal length with the exception of the last one. In this case, the last shipment-consolidation-cycle is the longest one.

ii) It is never optimal to use more than one cargo container (e.g., release more than one truckload) for an outbound delivery except for the outbound delivery made at the end of the last shipment-consolidation-cycle associated with a replenishment-cycle.

The first part of Property 1.5 relies on the fact that there is no inventory at the vendor in the last consolidation cycle. Hence, no inventory holding costs accumulate during this last cycle allowing us to lengthen the duration of consolidation. The second part part is due to the fact that it does not make sense to hold inventory in stock as long as we accumulate sufficient demand and have sufficient inventory to release a full truck-load.

5.3.4 A Comparison of Approximate and Exact Policies.

Numerical evidence shows that, in general, the optimal T_2 may be substantially larger than the optimal T_1. Hence, use of unequal shipment-consolidation-cycles seem to translate into cost savings for both the capacitated and uncapacitated problems. Exceptions are those problems where the cargo capacity constraints are not binding and the w/h ratio is high so that customer waiting is excessively expensive.

5.3.5 Future Research in Deterministic Models.

As we have seen, constant demand rate problems (with or without cargo capacity constraints) can be solved using nonlinear programming. When cargo capacity constraints are modeled explicitly, these are non-convex minimization problems because of the structure of the cargo costs. This is also true for common-carriage problems, which remain an area for future research, along with multiple market area and multi-item problems under cargo capacity constraints. Research on constant demand rate problems provides a foundation for analytical work on dynamic (time-varying deterministic) as well as stochastic problems.

5.4 Practical Coordinated/Integrated Policies for Stochastic Problems

Obviously, the simplistic policy discussed in the previous section is not appropriate in a probabilistic environment. For this more general case, the following policies can be adopted:

- (s, S, T_c) policy under which a delivery truck is dispatched every T_c time units, and inventory replenishment decisions are made following an (s, S) policy where s is the reorder level and S is the order-up-to level.

- (s, S, Q_c) policy under which a dispatch decision is taken when the consolidated weight exceeds Q_c, and inventory replenishment decisions are made following an (s, S) policy.

- $(s, S, \min\{T_c, S_{N_2(Q_c)}\})$ policy under which replenishment decisions follow an (s, S) policy and dispatch decisions follow a hybrid policy with parameters T_c and Q_c (i.e., we attempt to accumulate Q_c, but dispatch by time T_c if a consolidated load of Q_c has not been attained earlier).

The analysis of pure consolidation strategies (Section 5.1) provides a basis for computing optimal (s, S, T_c) and (s, S, Q_c) policies.

5.4.1 Integrated-Time-Based Dispatch Policy.

Let us consider the cost structure introduced in Section 5.3 (i.e., the case of a private-carriage without cargo capacity constraints) and discuss some results from Çetinkaya and Lee (2000) for computing the policy parameters for (s, S, T_c) policies.

By assuming that demands arrive according to a Poisson process, we can obtain an analytical expression for the expected long-run average cost function under the (s, S, T_c) policy using renewal theory. It is worth noting that, under the time-based dispatch policy, the inventory system under consideration can be treated as a periodic review system (Axsäter (2001)). That is, the inventory system is reviewed only in connection with periodic shipments to customers at which time the inventory position before a possible order is defined as the inventory on hand minus the demand that has been consolidated during the period. Consider the case where the replenishment lead time is negligible and an (s, S) policy with $s = -1$ and $S \geq 0$ is employed for replenishing the inventory. This replenishment rule guarantees that consolidated demand during a period can always be dispatched at the end of the period and on-hand inventory is always in the interval $[0, S]$ (see Axsäter (2001)). With this periodic review treatment, along with the standard inventory replenishment and holding costs, i.e., K_r and h, there is also a fixed cost K_c for dispatching and a customer waiting cost w per unit per unit of time. The integrated problem involves computing the order-up-to level, S, and the dispatch frequency, T_c, simultaneously.

Again, $C(S, T_c)$denotes the expected long-run average cost function. Using the renewal reward theorem, we have

$$C(S, T_c) = \frac{E[\text{Cost of an inventory-replenishment-cycle}]}{E[Y]T_c},$$

where Y denotes the number of shipment-consolidation-cycles within the replenishment-cycle. The main challenge in computing an analytical expression for $C(S, T_c)$ is in the calculation of $E[Y]$. Since demands arrive according to a Poisson process, say with parameter λ, then the sizes of the consolidated loads in successive shipment-consolidation-cycles are i.i.d. according to a Poisson distribution with parameter λT_c. This observation is useful in proving the following theorem.

THEOREM 1.6 *If demands arrive according to a Poisson process with parameter λ, then a continuous approximation for Y is provided by an*

Erlang random variable with scale parameter λT_c and shape parameter S. Thus,

$$E[Y] \approx \frac{S+1}{\lambda T_c}.$$

Using Theorem 1.6 and other results from the theory of stochastic processes, the cost function can be approximated by

$$C(S,T_c) \approx C_{app}(S,T_c) \equiv \frac{K_r\lambda}{S+1} + c_r\lambda + \frac{h\lambda T_c S}{S+1} + \frac{hS}{2} + \frac{K_c}{T_c} + \frac{w\lambda T_c}{2}.$$

Although $C_{app}(S,T_c)$is not jointly convex in S and T_c,it is easy to show that it has a unique minimizer.

THEOREM 1.7 *Provided that*

$$-\sqrt{\frac{2K_c\lambda}{w}} \leq \frac{K_r\lambda}{h} - \frac{1}{2},$$

the minimizer of$C_{app}(S,T_c)$ is the unique solution of

$$S = \sqrt{\frac{2K_r\lambda}{h} - 2\lambda T_c} - 1 \quad and \quad T_c = \sqrt{\frac{2K_c(S+1)}{\lambda[2hS + w(S+1)]}}.$$

Otherwise, the minimizer is at

$$S = 0 \quad and \quad T_c = \sqrt{\frac{2K_c}{\lambda w}}.$$

The minimizer of$C_{app}(S,T_c)$ gives an approximate solution for the problem of interest here. The exact optimal solution requires more effort but can be computed *numerically* using the approach developed in a recent paper by Axsäter (2001).

An important question that has been addressed in Axsäter (2001) is under what conditions the approximation given in Theorem 1.6 provides a good estimate of the exact value of $E[Y]$. In fact, the approximation may be poor if $S+1 \leq \lambda T_c$. However, it can be easily shown that if $S+1 \leq \lambda T_c$, then, by definition, $E[Y] = 1$ (Çetinkaya, Mutlu, and Lee (2002)). Consequently, a better approximation of $E[Y]$ is given by

$$E[Y] \approx \begin{cases} 1, & S+1 \leq \lambda T_C, \\ \frac{S+1}{\lambda T_c}, & S+1 > \lambda T_C. \end{cases}$$

Extensive numerical experiments illustrate that the performance of the above approximation is remarkably good (Çetinkaya, Mutlu, and Lee (2002)).

A significant amount of the classical inventory literature assumes that demands are satisfied as they arrive. That is, the traditional models focus on immediate outbound delivery policies where an outbound shipment has to be released each time a demand is received. Hence, in the existing literature, parameter K_c is treated as a sunk cost, and shipment consolidation opportunities are not modeled explicitly. As we have already mentioned, one of the uses of the integrated policy with shipment consolidation considerations is to provide a point of comparison between the operating costs under an immediate delivery policy and the integrated policies proposed here. This is helpful in both distribution policy and distribution system design. If demands are delivered as they arrive (without consolidation), then no waiting costs accumulate. In this case, our problem reduces to a pure inventory problem for which the order-up-to level S should be computed. For the case of Poisson demand arrivals with rate λ, the expected long-run average cost under an immediate delivery policy is given by (see Bhat (1984), pp. 435-436)

$$\frac{K_r\lambda}{S} + c_r\lambda + \frac{h(S+1)}{2} + K_c\lambda.$$

Thus, the optimal S value is again given by the standard EOQ formula. Using these results and comparing the cost of an optimal $(-1, S, T_c)$ policy with the cost of an immediate delivery policy, it is straightforward to identify the cases under which the optimal $(-1, S, T_c)$ policy outperforms an immediate delivery policy.

5.4.2 Integrated-Quantity-Based Dispatch Policy. Under the assumption that demands form a compound renewal process, some results have already been presented for computing the parameters for $(-1, S, Q_c)$ policies in Çetinkaya, Tekin, and Lee (2000). Under this policy, if $S < Q_c$, then the number of consolidation cycles within a replenishment-cycle, Y, is equal to 1, and, otherwise, Y is a random variable. Using the renewal reward theorem, in Çetinkaya, Tekin, and

Lee (2000) the cost function is given by

$$\mathcal{C}(S,Q_c) = \begin{cases} \frac{K_r+K_c}{E[X_k][M_G(Q_c)+1]} + \frac{c_r E[W_k]}{E[X_k]} \\ + \frac{w[Q_c M_G(Q_c) - \int_0^{Q_c}(Q_c-y)dM_G(y)]}{M_G(Q_c)+1} + hS, & \text{if } S < Q_c, \\ \\ \frac{K_r+K_c E[Y]}{E[X_k][M_G(Q_c)+1]E[Y]} + \frac{c_r E[W_k]}{E[X_k]} \\ + \frac{w[Q_c M_G(Q_c) - \int_0^{Q_c}(Q_c-y)dM_G(y)]}{M_G(Q_c)+1} \\ +hS - \frac{hE[W_k][M_G(Q_c)+1]E[Y(Y-1)]}{2E[Y]}, & \text{if } S \geq Q_c, \end{cases}$$

where $M_G(Q_c) \equiv E[N_2(Q_c)]$, i.e., $M_G(\cdot)$ denotes the renewal function associated with distribution function $G(\cdot)$ of W_k. *However, the Poisson arrivals assumption is no longer needed to obtain the above expression for the cost function.* Again, the main challenge in computing an analytical expression for the expected long-run average cost function is in the calculation of $E[Y]$ although another challenge also exists in the calculation of $E[Y(Y-1)]$.

Exact expressions of $E[Y]$ and $E[Y(Y-1)]$ cannot be represented in closed form even for the simpler cases of Poisson demand arrivals with exponential or uniform order weights. The following theorems provide approximations for $E[Y]$ and $E[Y(Y-1)]$.

THEOREM 1.8 *If the demand arrivals form a compound renewal process, then*

$$\begin{aligned} E[Y] &\approx 1 + \frac{S}{v_1} + \frac{v_2 - 2v_1^2}{2v_1^2}, \\ E[Y(Y-1)] &\approx \frac{S^2}{v_1^2} + \frac{2}{v_1^2}\left(\frac{v_2}{v_1} - v_1\right)S + \frac{2}{v_1^2}\left(-\frac{v_2}{2} + \frac{3v_2^2}{4v_1^2} - \frac{v_3}{3v_1}\right), \end{aligned}$$

where v_i, $i = 1,2,3$, denotes the i^{th} moment of $D_{N_2(Q_c)+1}$.

Note that the above approximations work for more general processes than the approximation of $E[Y]$ for the integrated-time-based dispatch policy, which is applicable only for Poisson demand processes (see Theorem 1.6). These approximations rely on the fact that

$$E[Y] = \sum_{k=1}^{\infty} P(Y \geq k) = \sum_{k=1}^{\infty} H^{(k-1)}(S) = 1 + M_H(S)$$

where $H(\cdot)$ denotes the distribution function of $D_{N_2(Q_c)+1}$, $H^{(k)}(\cdot)$ denotes the k-fold convolution of $H(\cdot)$, and $M_H(\cdot)$ denotes the renewal function associated with distribution function $H(\cdot)$. Hence, using the well-known results about asymptotic approximations for renewal functions, it is straightforward to validate the approximations in Theorem 1.8.

Using the approximate expressions of $E[Y]$ and $E[Y^2]$, we obtain an explicit approximate expression for $C(S, Q_c)$, denoted $C_{app}(S, Q_c)$, in terms of S and Q_c. Since this resulting expression is messy, it is not straightforward to make general conclusions about the convexity properties of $C_{app}(S, Q_c)$. In fact, $C_{app}(S, Q_c)$ may, or may not, be jointly convex in its variables depending on the properties of$G(\cdot)$, $M_G(\cdot)$, $H(\cdot)$, and the values of the model parameters. As a result, there are additional challenges to computing a global minimizer. Using an exhaustive search algorithm in two dimensional space, a detailed computational study examines the sensitivity of the optimal solution with respect to model parameters in Çetinkaya, Tekin, and Lee (2000). In particular, Çetinkaya, Tekin, and Lee (2000) reports several managerial insights regarding

1) the problem instances where it is preferable to consolidate demands
2) the sensitivity of optimal S and Q_c values to system parameters,
3) the potential savings that can be obtained by consolidating demands, and tradeoffs between waiting time induced by consolidating demands and costs saved.

Next, we discuss some of these insights.

Forms of the Optimal Policies The approximately optimal S and Q_c values, denoted by S^* and Q_c^*, respectively, imply one of the following three forms for the integrated policy under consideration:

- Form I. There is a single shipment consolidation cycle within a replenishment cycle. This is the case if $S^* = 0$, and, hence, $Q_c^* > S^*$.
- Form II. There are multiple consolidation cycles within a replenishment cycle. This is the case when $S^* > Q_c^* > 0$.
- Form III. Shipment consolidation does not make economic sense; hence $Q_c^* = 0$ whereas $S^* \geq 0$.

We say that shipment consolidation is a viable alternative if the resulting policy is of Form I or Form II. On the other hand, if the resulting policy is of Form III, then an immediate outbound dispatch policy is preferable because shipment consolidation does not make economic sense. We also note that if the policy is of Form I, then no inventory is held at the vendor's warehouse, i.e., the vendor's warehouse is operated as a transshipment point for consolidating orders. Thus, the proposed VMI arrangement is economically viable only if the optimal policy is of Form II.

The fundamental insight regarding the policy form can be summarized as follows: If $w \leq h$, then the optimal policy is of Form I. On the other hand, if $w > h$, then the structure of the policy (e.g., Form I, Form II, or Form III) depends on the ratios w/h and K_r/K_c as well as the arrival rate of the demand process.

In fact, for every problem instance considered in Çetinkaya, Tekin, and Lee (2000), there exist minimum and maximum threshold ratios for w/h after which the corresponding optimal policy changes from Form I to Form II and from Form II to Form III, respectively. These minimum and maximum threshold ratios define a region of w and h values over which the corresponding optimal policy is of Form II. Similarly, there also exist minimum and maximum threshold ratios for K_r/K_c after which the policy changes from Form I to Form II and from Form II to Form III, respectively.

Sensitivity of Q_c^* and S^* to Cost Parameters As we have already mentioned, the proposed VMI arrangement makes sense only if the optimal policy is of Form II where the vendor holds inventory and consolidates orders, simultaneously. Considering this particular case, numerical evidence in Çetinkaya, Tekin, and Lee (2000) suggests that, if the proposed VMI arrangement is viable, then:

- Increasing K_r has no substantial effect on Q_c^*, but results in an increase in S*. On the other hand, increasing K_c leads to a decrease in Q_c^* and an increase in S^*, simultaneously. As K_c increases, in an attempt to decrease shipment frequency, Q_c^* also increases which in turn results in an increase in average waiting costs. This increase in waiting costs is then compensated for through a decrease in holding costs via lowering 5*.

- An increase in h leads to a decrease in S^* but has almost no effect on Q_c^*. However, as w increases, S^* increases and Q_c^* decreases, simultaneously.

- In general, the effects of individual cost parameters on the Q_c^* and S* values seem to be complicated. This is mainly because these effects depend on the form of the optimal policy, e.g., Form I, Form II. or Form III.

- Although we do not have any analytical results regarding the shape of the cost function, numerical evidence suggests that $C(S^*, Q_c)$ is an EOQ-type convex function of Q_c. Similarly, $C(S, Q_c^*)$ seems to be an EOQ-type convex function of S. As a result, some of the intuitive EOQ-type observations still hold, e.g., the expected long-run average replenishment cost seems to be a decreasing convex function of 5 whereas the long-run average holding cost seems to be an increasing linear function of Q_c.

Cost Savings and Tradeoffs As we have mentioned earlier, if demands are delivered as they arrive (without consolidation), then no waiting costs accumulate. For the case of general demand processes considered here, the expected long-run average cost, denoted $\bar{C}(S)$ for the immediate delivery case, is given by

$$\begin{aligned}\bar{C}(S) = {} & \frac{K_r}{E[X_k][M_G(S)+1]} + \frac{c_r E[W_k] + K_c}{E[X_k]} + hS \\ & - \frac{hE[W_k](E[N_2(S)^2] - M_G(S))}{2[M_G(S+1]}\end{aligned} \tag{1.1}$$

Using the results presented so far and comparing the costs of optimal $(-1, S, T_c)$ and $(-1, S, Q_c)$ policies with the cost of an immediate delivery policy given by (1.1), we can obtain insightful *approximate* results relevant to distribution system and policy design. Exact results can also be obtained for some special cases of the problem as discussed in Section 5.4.3.

The numerical results presented in Çetinkaya, Tekin, and Lee (2000) demonstrate that the estimated cost savings due to consolidation depends on the values of the model parameters. Naturally, as K_c decreases, these cost savings diminish. Similarly, as w increases, $Q_c^* \to 0$ so that the estimated cost saving decreases. In fact, for slow-moving items where individual order weights are large enough to realize scale economies in outbound transportation, an immediate delivery policy (i.e., a policy of Form III) may be the most effective. On the other hand, for faster moving items with smaller individual order weights, the potential cost savings may be significant. In general, the costs saved critically depend on the w/h and K_r/K_c ratios as well as the magnitude of K_c and w.

Since the estimated cost saving is a result of the waiting time implied by the consolidation practice, it is important to analyze the *tradeoff between costs saved* and *timely delivery.* For implementation purposes, we can measure the timeliness of deliveries by the expected time between deliveries. Also, in evaluating the performance of a policy, we can consider the following two criteria simultaneously: i) the long-run average cost, and ii) the expected time between deliveries. According to our first criterion, the best policy is given by (Q_c^*, S^*) whereas this policy may not be as favorable according to our second criterion. In fact, under the second criterion, the best policy is the immediate delivery policy simply because "the smaller the Q_c, the higher the delivery frequency." Note that, for a given parameter set, implementing a policy with parameters (Q_c, S^*) where $Q_c > Q_c^*$ does not make sense because such a policy is dominated by (Q_c^*, S^*) according to *both* of the criteria under consideration.

Before concluding this section, it is worth emphasizing that the cost given by the minimum of $C_{app}(S, Q_c)$ is only an "estimate." The exact cost function requires the exact expression for the renewal function $M_H(S)$ (recall that $E[Y] = 1 + M_H(S)$). The approximations in Theorem 1.8 are based on the asymptotic approximations of $M_H(S)$. This seems to be a reasonable resolution since $M_H(S)$ may not have an easily computable form depending on the underlying order weight distribution (in fact, this is the case even under the assumption of exponential order weights). Hence, computational effort must be based on numerical methods and approximations for a wide class of distributions of practical interest. cSahin (1989) (pp. 41–53) gives an excellent overview of distributions for which the corresponding renewal functions do not assume an easy-to-compute form and presents results about the accuracy of asymptotic approximations. Typically, convergence to these approximations is quite fast. Next, we discuss a special case where it is easy to obtain an analytical expression for the exact cost function and an optimal solution for the integrated-quantity-based dispatch model.

5.4.3 A Simplified Integrated-Quantity-Based Dispatch Model for Unit Demand Arrivals.

Let us consider the case where the demands are of identical sizes. That is, each order demands one unit of the item or each demand weighs one unit. Observe that, unlike in the cases of the time-based dispatch or quantity-based dispatch with bulk demands, the number of shipment-consolidation-cycles within a replenishment-cycle, *Y*, is no longer a random variable after we fix *S* and Q_c (see Çetinkaya, Mutlu, and Lee (2002)). That is, *Y* is a constant, given by, say n, as in the case of the deterministic model in Section 5.3. One can substitute $S = (n-1)Q_c$, where n is an integer denoting the number of dispatch cycles within an inventory replenishment cycle, so that there is no inventory at the vendor during the last dispatch cycle of an inventory replenishment cycle. That is, the maximum inventory at the vendor's warehouse is $(n-1)Q_c$ whereas the order quantity is nQ_c. Hence, the expected long-run average cost per unit-time, denoted by $C(n, Q_c)$, is given by

$$C(n, Q_c) = c_r\lambda + \frac{K_r\lambda}{nQ_c} + \frac{K_c\lambda}{Q_c} + \frac{h(n-1)Q_c}{2} + \frac{w(Q_c-1)}{2}.$$

The minimizer of this function can be obtained in a way that is very similar to the approximate solution of the deterministic model in Section 5.3. Again, let n^*, and Q_c^* denote the optimal n and Q_c values, respectively. Again, we define

$$n_0 = \sqrt{\frac{K_r(w-h)}{K_c h}}.$$

It can be easily shown that

- If $w \leq h$, then $n^* = 1$ and $Q_c{}^*$ is either $\left\lfloor \sqrt{2(K_r + K_c)\lambda/w} \right\rfloor$ or $\left\lceil \sqrt{2(K_r + K_c)\lambda/w} \right\rceil$ depending on which one yields a lower value of $C(1, Q_c)$.
- If $w > h$, then the optimal solution is given by

  $$\min\left\{C(n_1, q_1), C(n_1, q_2), C(n_2, q_3), C(n_2, q_4)\right\}$$

  where

  $$n_1 = \lfloor n_0 \rfloor, n_2 = \lceil n_0 \rceil, q_1 = \lfloor q(n_1) \rfloor, q_2 = \lceil q(n_1) \rceil,$$

  $$q_3 = \lfloor q(n_2) \rfloor, q_4 = \lceil q(n_2) \rceil, \quad \text{and}$$

  $$q(n) = \sqrt{\frac{2(K_r + nK_c)\lambda}{n[w + h(n-1)]}}.$$

Once n^* and $Q_c{}^*$ are computed, the optimal S, denoted by S^*, is given by $(n^*-1)Q_c{}^*$. Of course, the above *stationary* policy, specified by (n^*, Q_c^*), can be easily improved in the manner discussed in Section 5.3.2. We revisit the issue of suboptimality of stationary policies in Section 5.4.6.

5.4.4 Comparison of Practical Integrated Policies for Poisson Demand. Now, in order to obtain a point of *exact* comparison between the integrated-quantity-based dispatch model and the integrated-time-based dispatch model, suppose that the demand arrivals form a pure (unit) Poisson process with parameter λ. Under this assumption, Çetinkaya, Mutlu, and Lee (2002) present analytical and numerical results. This work shows that the cost savings obtained through integrated-quantity-based and inventory/hybrid dispatch policies can be substantial relative to the integrated-time-based dispatch policies. The cost savings obtained by integrated-hybrid dispatch policies are estimated through simulation, whereas the exact costs of the optimal integrated-quantity-based and optimal inventory/time-based dispatch policies are computed using the exact approaches presented in Çetinkaya, Mutlu, and Lee (2002) and Axsäter (2001), respectively.

5.4.5 Extensions for the Coordinated/Integrated Stochastic Models.

Extensions for the Integrated- Time-Based Dispatch Model The consideration of specific arrival patterns, such as Erlang processes, is of practical importance. For such general problems, the resulting inventory and consolidation processes are non-Markovian. By employing the "method of stages" (Kleinrock (1975), p. 119–126), these processes can be transformed into Markovian processes at the cost of enlarging the state space. Thus, development of efficient computational procedures for such arrival processes is an area for future investigation along with sophisticated simulation models that investigate timely delivery versus cost tradeoffs using practical data.

The integrated-time-based dispatch model presented here ignores the issue of replenishment lead time uncertainty. Given the popularity and success of just-in-time manufacturing in some industries, replenishment lead time uncertainty may not be a critical problem for practical purposes. However, it remains a challenging theoretical problem. Other natural extensions include the case where the vendor distributes multiple items to multiple markets, common-carrier problems, and finite cargo capacity problems.

Extensions for the Integrated-Quantity-Based Dispatch Model Again, natural extensions include those of the integrated-time-based dispatch model. A challenging extension (of both the integrated-time-based and the inventory/quantity-based dispatch models) is the computation of the parameters of an integrated-hybrid-based dispatch policy. Under such a policy, the sizes of consolidated loads in successive shipment-consolidation-cycles are not i.i.d. in general. Thus, computing $E[Y]$, as well as the distribution of Y, is an important challenge. Some special cases of the problem may be modeled using Markov renewal theory and stochastic dynamic programming. However, development of a sophisticated simulation model is also an important research endeavor, in particular for model validation purposes in practice.

The integrated stochastic models discussed so far provide basic tools for a comparative analysis of immediate delivery, TDD, and FTL (load-optimized delivery) provisions. However, some fine tuning of these models is required for full scale practical application. For example, both models overlook the impact of truck/cargo capacity considerations whereas, under the assumption of FTL provisions, the dispatch quantity is influenced by the truck/cargo capacity and the corresponding cost structure. Incorporation of capacity restrictions in these models introduces challenging stochastic optimization problems especially for the case of bulk arrival (i.e., compound renewal) processes. If cargo capacity constraints are ignored, performing a comparative analysis of immediate delivery, TDD, and FTL provisions is straightforward. In this case, the cost of an optimal immediate delivery (i.e., an LTL shipment) policy is obtained simply by minimizing $\bar{C}(S)$ in Equation (1.1), whereas the cost of an optimal TDD policy is given by the results of the integrated-time-based dispatch model. The optimal cost under FTL provisions can be computed simply by setting Q_c equal to the size of an FTL shipment in the integrated-time-based dispatch model.

5.4.6 Exact Policies. As we have illustrated in Section 5.3 and mentioned in Section 5.4.3, the stationary policies discussed in this chapter may be suboptimal in the class of all feasible policies. Nevertheless, they are easy to implement and compute in practical situations. To characterize *exact* optimal policies, we can utilize dynamic programming. This topic remains a challenging area for future investigation.

For illustrative purposes, consider a variation of the problem described in the previous section. Suppose that customers place their orders on a periodic basis, e.g., at the beginning of each week. Now, with a slight change of notation, let $W_1, W_2, \ldots$ denote the aggregate demands from customers located in a given market area in successive periods indexed by $k = 1, 2, \ldots$. Ideally, $W_k, k = 1, 2, \ldots$ should be shipped immediately after the order is placed. However, we consider the case where the vendor has the liberty of not delivering small orders until an economical dispatch quantity accumulates. As we have mentioned previously, such an action can only be taken at the expense of customer waiting costs and inventory carrying costs. Waiting costs represent an opportunity loss in delayed receipt of revenue as well as a goodwill penalty. Although customers are willing to wait during a negotiable delivery *time-window,* acceptable service should be guaranteed by imposing a maximum waiting time for each order. A time-window is a grace period for delivery timing, a concept from the vehicle routing literature applicable in the context of VMI, 3PW/D, and TDD agreements. Let $\delta_1, \delta_2, \ldots$ denote the maximum number of periods that the vendor can delay the delivery of $W_1, W_2, \ldots$ respectively. That is, a shipment of size W_k is expected at the beginning of period k, but the vendor may delay this delivery until time t such that $k \leq t \leq k + \delta_k$. Hence, interval $[k, k + \delta_k]$ is called the *time-window* for $W_k, k = 1, 2, \ldots$.

If we ignore the outbound transportation costs, assume that all time-windows are of length zero, and consider time varying deterministic demands, the above problem reduces to the well known Wagner-Whitin lot-sizing problem (see Aggarwal and Park (1993); Federgruen and Tzur (1991); Wagelmans, Van Hoesel, and Kolen (1992); Wagner and Whitin (1958); Zangwill (1966) for meritorious studies on the Wagner-Whitin problem). Nevertheless, if transportation costs matter and consolidation is a feasible alternative (i.e., time-windows are nonzero), then the problem is challenging. The crucial questions that must be answered are: i) how often and in what quantities to replenish stock so that replenishment, holding, and waiting costs are minimized, ii) how often to dispatch a vehicle so that timely service requirements are satisfied, and iii) how large the dispatch-quantity should be so that transportation scale economies are not sacrificed. Furthermore, in answering these questions, the inventory replenishment policy at the vendor's warehouse and the optimal dispatch quantities/release times for shipments to the customer should be computed simultaneously. The case of time varying deterministic demands, where time-window considerations are modeled but outbound transportation costs are ignored, is studied in Lee, Çetinkaya, and Wagelmans (2001). Also, the case of time varying deterministic demands where time-window considerations are ignored but outbound cargo costs and capacity are considered explicitly, is studied in Lee, Çetinkaya, and Jaruphongsa (2002). The general problem with either deterministic or stochastic demands remains an area for future investigation. As we have mentioned, this class of problems is solved using dynamic programming based algorithms. For example, Lee, Çetinkaya, and Jaruphongsa (2002) and Lee, Çetinkaya, and Wagelmans (2001) develop polynomial time optimal dynamic programming algorithms. Although DP is a conceptually powerful technique, its computational limitations stress the importance of analytical results regarding the structure of the optimal solution. Incorporation of time-windows, explicit consideration of general freight cost structures, and cargo capacity constraints increase the computational requirements for this class of problems.

6. Conclusion

This chapter provides a review of analytical models for a general class of coordination problems applicable in the context of VMI, 3PW/D and TDD practices, and it introduces new avenues for research in the theory of supply-chain management. By developing a theoretical framework for understanding and justifying integrated inventory and outbound shipment consolidation practices, research in this area may have a significant impact on some of the many challenging practical problems in supply-chain management.

Acknowledgements

This research was supported by NSF Grant CAREER/DMI-0093654.

References

Axsäter, S. (2000), *Inventory Control,* Kluwer, Boston.

Axsäter, S. (2001), "A Note on "Stock Replenishment and Shipment Scheduling for Vendor-Managed Inventory Systems," *Management Science,* 47-9, pp. 1306–1310.

Abdelwahab, W.M. and M. Sargious (1990), "Freight Rate Structure and Optimal Shipment Size in Freight Transportation," *Logistics and Transportation Review,* 6-3, pp. 271–292.

Aggarwal, A. and J.K. Park (1993), "Improved Algorithms for Economic Lot-Size Problems," *Operations Research,* 41, pp. 549–571.

Anily, S. and A. Federgruen (1990), "One Warehouse Multiple Retailer Systems with Vehicle Routing Costs," *Management Science,* 36, pp. 92–114.

Anily, S. and A. Federgruen (1993), "Two-Echelon Distribution Systems with Vehicle Routing Costs and Central Inventories," *Operations Research,* 41, pp. 37–47.

Arcelus, F.J. and J.E. Rowcroft (1991), "Small Order Transportation Costs in Inventory Control," *Logistics and Transportation Review,* 27-1, pp. 3–13.

Arcelus, F.J. and J.E. Rowcroft (1993), "Freight Rates for Small Shipments," *International Journal of Production Economics,* 30–31, pp. 571–577.

Arntzen, B.C., G.G. Brown, T.P. Harrison, and L.L. Trafton (1995), "Global Supply-Chain Management at Digital Corporation," *Interfaces,* 25-1, pp. 69–93.

Aucamp, D.C. (1982), "Nonlinear Freight Costs in the EOQ Problem," *European Journal of Operational Research,* 9, pp. 61–62.

Aviv, Y. and A. Federgruen (1998), "The Operational Benefits of Information Sharing and Vendor Managed Inventory Programs," Technical Report, Department of IE and OR, Columbia University.

Ballou, R.H. (1999), *Business Logistics Management,* Fourth Edition, Prentice Hall.

Banerjee, A. (1986), "On a Quantity Discount Pricing Model to Increase Vendor Profits," *Management Science,* 32, pp. 1513–1517.

Banerjee, A. (1986), "A Joint Economic-Lot-Size Model for Purchaser And Vendor," *Decision Sciences,* 17, pp. 292–311.

Banerjee, A. and S. Burton (1994), "Coordinated vs. Independent Inventory Replenishment Policies for a Vendor and Multiple Buyers," *International Journal of Production Economics,* 35, pp. 215–222.

Baumol, W.J. and H.D. Vinod (1970), "An Inventory Theoretic Model of Freight Transport Demand," *Management Science,* 16, pp. 413–421.

Bhat U.N., *Elements of Applied Stochastic Processes,* John Wiley and Sons, 1984, New York.

Blumenfeld, D.E., L.D. Burns, C.F. Daganzo, M.C. Frick, and R.W. Hall (1987), "Reducing Logistics Costs at General Motors," *Interfaces,* 17-1, pp. 26–47.

Blumenfeld, D.E., L.D. Burns, J.D. Diltz and C.F. Daganzo (1985), "Analyzing Tradeoffs Between Transportation, Inventory, and Production Costs on Freight Networks," *Transportation Research,* 19B-5, pp. 361–380.

Boyaci, T. and G. Gallego (2002), "Coordination Issues in a Simple Supply-Chain," *International Journal of Production Economics,* 77, pp.95–11.

Bramel, J. and D. Simchi-Levi (1996), "Probabilistic Analyses and Practical Algorithms for the Vehicle Routing Problem with Time Windows," *Operations Research,* 44-3, pp. 501–509.

Bramel, J. and D. Simchi-Levi (1997), *The Logic of Logistics,* Springer-Verlag, New York.

Brennan, J.J. (1981), *Models and Analysis of Temporal Consolidation,* Doctoral Dissertation, University of California at Los Angeles.

Bookbinder, J.H. and J.K. Higginson (2002), "Probabilistic Modeling of Freight Consolidation by Private Carriage," *Transportation Research E,* 38-5, pp. 305–318.

Bourland, K., S. Powell, and D. Pyke (1996), "Exploiting Timely Demand Information to Reduce Inventories," *European Journal of Operational Research,* 92, pp. 239–253.

Bowersox, D.J. (1978), *Logistical Management,* Second Edition, Macmillan Publishing, New York.

Burns, L.D., R.W. Hall, D.E. Blumenfeld and C.F. Daganzo (1985), "Distribution Strategies that Minimize Transportation and Inventory Costs," *Operations Research,* 33-3, pp. 469–490.

Campbell, J.F. (1990), "Designing Logistics Systems by Analyzing Transportation, Inventory and Terminal Cost Tradeoffs," *Journal of Business Logistics,* 11-1, pp. 159–179.

Carter, J.R. and B.G. Ferrin (1996), "Transportation Costs and Inventory Management: Why Transportation Costs Matter," *Production and Inventory Management Journal,* 3rd Quarter, pp. 58–62.

Carter J.R., B.G. Ferrin, and C.R. Carter (1995), "The Effect of Less-Than-Truckload Rates on the Purchase Order Lot-Size Decisions," *Transportation Journal,* Spring, pp. 35–44.

Chan, L.M.A., A. Muriel, Z.-J.M. Shen, D. Simchi-Levi, C.P. Teo (2002), "Effective Zero-Inventory-Ordering Policies for the Single-Warehouse Multi-Retailer Problem with Piecewise Linear Cost Structures," *Management Science,* 48-11, pp. 1446-1460.

Constable, G.K. and D.C. Whybark, (1978) "The Interaction of Transportation and Inventory Decisions," *Decision Sciences,* 9, pp. 688–699.

Çetinkaya, S. and J. H. Bookbinder (2002), "Stochastic Models for the Dispatch of Consolidated Shipments," *Transportation Research B,* 37-8, pp. 747–768, 2003.

Çetinkaya, S. and Bookbinder, J.H. (1997), "Time-and-Quantity Policy for Consolidated Shipments," presented at the *Institute for Operations Research and the Management Sciences,* Dallas, TX, November 1997.

Çetinkaya, S. and C.-Y. Lee (2000), "Stock Replenishment and Shipment Scheduling for Vendor Managed Inventory Systems," *Management Science,* 46-2, pp. 217–232.

Çetinkaya, S. and C.-Y. Lee, (2002), "Optimal Outbound Dispatch Policies: Modeling Inventory and Cargo Capacity," *Naval Research Logistics,* 49-6, pp. 531–556.

Çetinkaya, S., F. Mutlu, and C.-Y. Lee (2002), "A Comparison of Outbound Dispatch Policies for Vendor-Managed Inventory Systems," Technical Report, Industrial Engineering Department, Texas A&M University.

Çetinkaya, S., E. Tekin, and C.-Y. Lee (2000), "A Stochastic VMI Model for Joint Inventory Replenishment and Shipment Release Decisions," Technical Report, Industrial Engineering Department, Texas A&M University. A synopsis of this paper has appeared in *MSOM Conference Proceedings,* Ann Arbor, Michigan, June, 2000.

Closs, D.J. and R.L. Cook (1987), "Multi-Stage Transportation Consolidation Analysis Using Dynamic Simulation," *International Journal of Physical Distribution and Materials Management,* 17-3, pp. 28–45.

Cooper, M.C. (1984), "Cost and Delivery Time Implications of Freight Consolidation and Warehouse Strategies," *International Journal of Physical Distribution and Materials Management,* 14-6, pp. 47–67.

Cottrill, K. (1997), "Reforging the Supply-Chain," *Journal of Business Strategy,* 18-6, pp. 35–39.

Daganzo, C.F. (1988), "Shipment Composition Enhancement at a Consolidation Center," *Transportation Research,* 22B-2, pp. 103–124.

Daganzo, C.F. (1996), *Logistics Systems Analysis,* Second Edition, Springer-Verlag, Berlin.

Davis, T. (1993), "Effective Supply-Chain Management," *Sloan Management Review,* Summer, pp. 35–46.

Diaby, M. and A. Martel (1993), "Dynamic Lot-Sizing for Multi-Echelon Distribution Systems with Purchasing and Transportation Price Discounts," *Operations Research,* 41, pp. 48–59.

Federgruen, A. and D. Simchi-Levi (1995), "Analytical Analysis of Vehicle Routing and Inventory Routing Problems," in M. Ball, T. Magnanti, C. Monma, and G. Nemhauser (eds.), *Handbooks on Operations Research and Management Science,* the volume on *Network Routing,* pp. 297–373.

Federgruen, A. and M. Tzur (1991), "A Simple Forward Algorithm to Solve General Dynamic Lot-Sizing Models with n Periods in $O(n \log n)$ or $O(n)$ Time," *Management Science,* 37, pp. 909–925.

Gallego, G. and D. Simchi-Levi (1990), "On the Effectiveness of Direct Shipping Strategy for the One-Warehouse Multi-Retailer R-Systems," *Management Science,* 36-2, pp. 240–243.

Gans, N. and G. van Ryzin (1999), "Dynamic Vehicle Dispatching: Optimal Heavy Traffic Performance and Practical Insights," *Operations Research,* 47, pp. 675–692.

Goyal, S.K. (1976), "An Integrated Inventory Model for a Single Supplier Single Customer Problem," *International Journal of Production Research,* 15, pp. 107–111.

Goyal, S.K. (1987), "Comment on:A Generalized Quantity Discount Pricing Model to Increase Supplier's Profits," *Management Science,* 33, pp. 1635–1636.

Goyal, S.K. and Y.P. Gupta (1989), "Integrated Inventory Models: The Buyer-Vendor Coordination," *European Journal of Operational Research,* 41, pp. 261–269.

Gupta, O.K. (1992), "A Lot-Size Model with Discrete Transportation Costs," *Computers and Industrial Engineering,* 22-4, pp. 397–402.

Gupta, Y.P. and P.K. Bagchi (1987), "Inbound Freight Consolidation under Just-In-Time Procurement: Application of Clearing Models," *Journal of Business Logistics,* 8-2, pp. 74–94.

Hadley, G. and T.M. Whitin (1963), *Analysis of Inventory Systems,* Prentice-Hall.

Hahm, J. and C. Yano (1992), "The Economic Lot and Delivery Scheduling Problem: The Single Item Case," *International Journal of Production Economics,* 28, pp. 235–252.

Hahm, J. and C. Yano (1995a), "The Economic Lot and Delivery Scheduling Problem: The Common Cycle Case," *IIE Transactions,* 27, pp. 113–125.

Hahm, J. and C. Yano (1995b), "The Economic Lot and Delivery Scheduling Problem: Models for Nested Schedules," *IIE Transactions,* 27, pp. 126–139.

Hall, R.W. (1987), "Consolidation Strategy: Inventory, Vehicles, and Terminals," *Journal of Business Logistics,* 8-2, pp. 57–73.

Hall, R.W. (1991), "Comments on One-Warehouse Multiple Retailer Systems on Vehicle Routing Costs," *Management Science,* 37, pp. 1496–1499.

Hall, R.W. and M. Racer (1995), "Transportation with Common Carrier and Private Fleets: System Assignment and Shipment Frequency Optimization," *IIE Transactions,* 27, pp. 217–225.

Henig, M., Y. Gerchak, R. Ernst, and D.F. Pyke (1997), "An Inventory Model Embedded In Designing a Supply Contract," *Management Science,* 43-2, pp. 184–189.

Higginson, J.K. (1995), "Recurrent Decision Approaches to Shipment Release Timing in Freight Consolidation," *International Journal of Physical Distribution and Logistics Management,* 25-5, pp. 3–23.

Higginson, J.K. and J.H. Bookbinder (1994), "Policy Recommendations for a Shipment Consolidation Program," *Journal of Business Logistics,* 15-1, pp. 87–112.

Higginson, J.K. and J.H. Bookbinder (1995), "Markovian Decision Processes in Shipment Consolidation," *Transportation Science,* 29-3, pp. 242–255.

Jackson, G.C. (1981), "Evaluating Order Consolidation Strategies Using Simulation," *Journal of Business Logistics,* 2-2, pp. 110–138.

Joglekar, P.N. (1988), "Comments on a Quantity Discount Pricing Model to Increase Vendor Profits," *Management Science,* 34, pp. 1391–1400.

Joglekar, P. and S. Tharthare (1990), "The Individually Responsible and Rational Decision Approach to Economic Lot-Sizes for One Vendor and Many Purchasers," *Decision Sciences,* 21, pp. 492–506.

Kleinrock, L. (1975), *Queueing Systems, Volume I: Theory,* John Wiley & Sons, New York.

Kleywegt, A., V. Nori, and M. Savelsberg (1998), “A Computational Approach for the Inventory Routing Problem,” Technical Report, Georgia Institute of Technology.

Knowles, T.W. and P. Pantumsinchai (1988), “All-Units Discounts for Standard Container Sizes,” *Decision Sciences,* 19, pp. 848–857.

Lee, C.-Y. (1986), “The Ecomonic Order Quantity for Freight Discount Costs,” *IIE Transactions,* 18-3, pp. 318–320.

Lee, C.-Y. (1989), “A Solution to the Multiple Set-Up Problem with Dynamic Demand,” *IIE Transactions,* 21-3, pp. 266–270.

Lee, C.-Y., S. Çetinkaya, W. Jaruphongsa (2002), “A Dynamic Model for Inventory Lot-Sizing and Outbound Shipment Scheduling at a Third Party Warehouse,” *Operations Research,* 51-5, pp. 735–747.

Lee, C.-Y., S. Çetinkaya, and A.P.M. Wagelmans (2001), “Dynamic Lot-Sizing Model with Demand Time Windows,” *Management Science,* 47-10, pp. 1384–1395.

Lee, H., and C. Billington (1992), “Managing Supply-Chain Inventory,” *Sloan Management Review,* Spring, pp. 65–73.

Lee, H.L. and M.J. Rosenblatt (1986), “A Generalized Quantity Discount Pricing Model to Increase Supplier’s Profits,” *Management Science,* 32, pp. 1177–1185.

Lee, H., V. Padmanabban, and S. Whang (1997), “Information Distortion in a Supply-Chain,” *Management Science,* 43-4, pp. 546–558.

Masters, J.M. (1980), “The Effects Of Freight Consolidation On Customer Service,” *Journal of Business Logistics,* 2-1, pp. 55–74.

Medhi, J. (1994), *Stochastic Processes,* Second Edition, Wiley Eastern Ltd., New Delhi.

Minkoff, A.S. (1993), “A Markov Decision Model and Decomposition Heuristic for Dynamic Vehicle Dispatching,” *Operations Research,* 41-1, pp. 77–90.

Newbourne, M.J. and C. Barrett (1972), “Freight Consolidation and the Shipper,” Parts 1 to 5, *Transportation and Distribution Management,* 12, Numbers 2 to 6.

Parker, K. (1996), “Demand Management and Beyond,” *Manufacturing Systems,* June, pp. 2A–14A.

Pollock, T. (1978), “A Management Guide to LTL Consolidation.” *Traffic World,* April 3, pp. 29–35.

Pooley, J. and A.J. Stenger (1992), “Modeling and Evaluating Shipment Consolidation in a Logistics System,” *Journal of Business Logistics,* 13-2, pp. 153–174.

Popken, D.A. (1994), “An Algorithm for the Multi-Attribute, Multi-Commodity Flow Problem with Freight Consolidation and Inventory Costs,” *Operations Research,* 42-2, pp. 274–286.

Powell, W.B. (1985), "Analysis of Vehicle Holding and Cancellation Strategies in Bulk Arrival, Bulk Service Queues," *Transportation Science,* 19-4, pp. 352–377.

Powell, W.B. and P. Humblet (1986), "The Bulk Service Queue with General Control Strategy: Theoretical Analysis and a New Computational Procedure," *Operations Research,* 34-2, pp. 267–275.

Russell, R.M. and L. Krajewski (1991), "Optimal Purchase and Transportation Cost Lot-Sizing for a Single Item" *Decision Sciences,* 22, pp. 940–952.

Şahin (1989), İ, *Regenerative Inventory Systems,* Bilkent University Lecture Series, Springer-Verlag, Ankara/Turkey.

Schenck, J. and Mclnerney, J. (1998), "Applying Vendor-Managed Inventory to the Apparel Industry," *Automatic I.D. News.,* 14-6, pp. 36–38, May.

Schwarz, L.B. (1973), "A Simple Continuous Review Deterministic One-Warehouse *N*-Retailer Inventory Problem," *Management Science,* 19, pp. 555–566.

Sethi, S.P. (1984), "A Quantity Discount Lot-Size Model with Disposals," *International Journal of Production Research,* 22, pp. 31–39.

Sheffi, Y., B. Eskandari, and H.N. Koutsopoulos (1988), "Transportation Mode Choice Based on Total Logistics Costs," *Journal of Business Logistics,* 9-2, pp. 137–154.

Smith, R.D., T.M. Corsi, and C.M. Grimm (1990), "Motor Carrier Strategies and Performance," *Transportation Research A,* 24A-3, pp. 201–210.

Stalk, G, P. Evans, and L.E. Shulman (1992), "Competing on Capabilities: The New Rules of Corporate Strategy," *Harvard Business Review,* March-April, pp. 57–69.

Stevens, G.C. (1989), "Integrating the Supply-Chain," *International Journal of Physical Distribution and Materials Management,* 8, pp. 3–8.

Stidham, S., Jr. (1977), "Cost Models for Stochastic Clearing Systems," *Operations Research,* 25-1, pp. 100–127.

Tayur, S., R. Ganeshan, and M. Magazine (1999), *Quantitative Models for Supply-Chain Management,* Kluwer, Boston.

Tersine, R.J. and S. Barman (1991), "Economic Inventory/Transport Lot-Sizing with Quantity and Freight Rate Discounts," *Decision Sciences,* 22-5, pp. 1171–1179.

Tyworth, J.E. (1992), "Modeling Transportation-Inventory Tradeoffs in a Stochastic Setting," *Journal of Business Logistics,* 12-2, pp. 97–124.

Van Eijs, M.J.G. (1994), "Multi-item Inventory Systems with Joint Ordering and Transportation Decisions," *International Journal of Production Economics,* 35, pp. 285–292.

Viswanathan, S. and K. Mathur (1997), "Integrating Routing and Inventory Decisions in One-Warehouse Multi-Retailer Multi-Product Distribution Systems," *Management Science,* 43-3, pp. 294–312.

Wagelmans, A., S. Van Hoesel, and A. Kolen (1992), "Economic Lot-Sizing: An $O(n \log n)$ Algorithm that Runs in Linear Time in the Wagner-Whitin Case," *Operations Research,* 40, (1992), pp. S145–S156.

Wagner, H.M. and T.M. Whitin (1958), "Dynamic Version of the Economic Lot-Size Model," *Management Science,* 5, pp.89–96.

Zangwill, W.I. (1966), "A Deterministic Multi-Period Production Scheduling Model with Backlogging," *Management Science,* 13, pp. 105–119.

This page intentionally left blank

Chapter 2

A NEAR-OPTIMAL ORDER-BASED INVENTORY ALLOCATION RULE IN AN ASSEMBLE-TO-ORDER SYSTEM AND ITS APPLICATIONS TO RESOURCE ALLOCATION PROBLEMS

Yalçın Akçay
College of Administrative Sciences and Economics
Koç University
Rumeli Feneri Yolu
Sariyer
Istanbul 34450
Turkey
yakcay@ku.edu.tr

Susan H. Xu
Department of Supply Chain and Information Systems
The Smeal College of Business Administration
Penn State University
University Park, Pennsylvania 16802
shx@psu.edu

Abstract Assemble-to-order (ATO) manufacturing strategy has taken over the more traditional make-to-stock (MTS) strategy in many high-tech firms. ATO strategy has enabled these firms to deliver customized demand timely and to benefit from risk pooling due to component commonality. However, multi-component, multi-product ATO systems pose challenging inventory management problems. In this chapter, we study the component allocation problem given a specific replenishment policy and realized customer demands. We model the problem as a general multidimensional knapsack problem (MDKP) and propose the primal effective capacity heuristic (PECH) as an effective and simple approximate solution procedure for this NP-hard problem. Although the heuristic is primarily designed for the component allocation problem in an ATO system, we suggest that it is a general solution method for a wide range of resource allocation problems. We demonstrate the effectiveness of the heuristic through an extensive computational study which covers problems from the literature as well as randomly generated instances of the general and 0-1 MDKP. In our study, we compare the performance of the heuristic with other approximate solution procedures from the ATO system and integer programming literature.

1. Introduction

In response to increasing pressure from customers for fast delivery, mass customization, and decreasing life cycles of products, many high-tech firms have adopted the assemble-to-order (ATO) in place of the more traditional make-to-stock (MTS) strategy. In contrast to MTS, which keeps inventory at the *end-product* level, ATO keeps inventory at the *component* level. When a customer order is received, the components required are pulled from inventory and the end-product is assembled and delivered to the customer. The ATO strategy is especially beneficial to firms with significant component replenishment lead times and negligible final assembly times. The ATO strategy postpones the point of commitment of components to specific products, and thus, increases the probability of meeting a customized demand timely and at low cost (Lee and Tang, 1997). Furthermore, by using common components and modules in the final assembly, ATO is better protected against demand variability due to risk pooling effect. By successfully implementing ATO, for example, the Dell Corporation has reduced inventory costs, mitigated the effect of product obsolescence, and thrived in the competitive PC market (Agrawal and Cohen, 2000). The IBM Personal Computing Systems Group has also successfully transformed from its previous MTS operation to the current ATO practice (Chen et al., 2000).

In this chapter, we consider a multi-product, multi-component, periodic-review ATO system that uses the independent base-stock (order-up-to level) policy for inventory replenishment. We assume that the replenishment lead time of each component is an integer multiple of the review interval and can be different for different components. Product demands in each period are integer-valued, possibly correlated, random variables, with each product being assembled from multiple units of a subset of components. The system quotes a pre-specified response time window for each product and receives a reward if the demand for that product is filled within its response time window, where an order is said to be filled only if all the components requested by the order are present.

The multi-component, multi-product ATO system poses challenging inventory management problems. One such problem is to determine inventory replenishment levels without full information on product demands. Another problem is to make component allocation decisions based on available component inventories and realized product demands. Because fulfilling a customer order requires simultaneous availability of multiple units of several components, the optimal component allocation decisions lead to an NP-hard combinatorial optimization problem. Although such a problem can be solved analytically, it is often impractical to implement the optimal policy due to computational complexity of the solution method and complex structure of the policy. This ATO inventory management problem can be modelled as a two-stage stochastic integer program. In the first stage, the optimal base-stock levels of various components for each review period are determined, subject to a given inventory investment budget constraint. This decision is made at the beginning of the first period without knowledge of product demands in subsequent periods. In the second stage, product demands are observed and component allocation decisions are made based on the inventory on-hand and the realized demands, such that the total reward of *filled* orders within their respective response time windows is maximized. Therefore, the optimal component replenishment policy depends on the component allocation rule to be applied after product demand realization, which makes the analytical solutions even more difficult to obtain. When the rewards of all product types are identical, our objective function reduces to the so-called aggregated *type-II* service level, also known as the *fill rate.* In the inventory literature, the *type-I* service level measures *the proportion of periods* in which the demand of a product (or the aggregated demand of all products) is met, whereas the type-II service level measures *the proportion of the demand* of a product (or the proportion of the aggregated demand of all products) that is satisfied. Indeed, a major disadvantage of the type-I service level is that it disregards the batch size effect; in contrast, the type-II service level often provides a much better picture of service from customers' perspective (Axsäter, 2000).

In this chapter, we focus on the second-stage problem of the two-stage model described above. A detailed analysis of the first-stage problem, in which the component base-stock levels are optimized, can be found in Akçay and Xu, 2002. As we noted earlier, the multi-component, multi-product component allocation problem is often a large-scale integer program and is computationally demanding. Indeed, we will show that our component allocation problem is a large-scale, *general multidimensional knapsack problem* (MDKP). The general MDKP fills a single knapsack subject to multiple resource constraints and permits multiple units of an item to be placed in the knapsack. The objective is to maximize the total value of the items placed in the knapsack. It has been applied to many different settings including capital budgeting, cargo loading, project selection, resource allocation, scheduling, and portfolio optimization. The MDKP is NP-hard, even with a single resource constraint $m = 1$ (Garey and Johnson, 1979). Although there exists an extensive literature addressing the solution procedure for the 0-1 MDKP, it appears that there are just a few solution procedures available to solve the large-scale, general MDKP (Lin, 1998). In this chapter, we shall propose a simple, yet effective, *order-based component allocation* rule that can be implemented in a real-time ATO environment. Unlike the component-based allocation rules in which the component allocation decisions are made independently across different components, such as the fixed priority (Zhang, 1997) and fair shares (Agrawal and Cohen, 2000) rules, our rule commits a component to an order only if it leads in the fulfillment of the order within the quoted time window; otherwise, the component is saved for other customer orders.

A common feature of the heuristics in the MDKP literature is the concept of the *effective gradient.* The effective gradient of an item is defined as the ratio (referred to as the "bang-for-buck ratio") of the reward of that item to its *aggregate* resource consumption among all resource constraints. The effective gradient method then increments the value of the item with the largest effective gradient in the solution. Loosely speaking, the effective gradient method and its variants use a "max-sum" criterion to make item selection decisions. In effect, these methods reduce a multidimensional problem to a single dimensional problem. We believe that the use of the *aggregate* resource consumption of an item in the effective gradient measure has several drawbacks. First, it does not distinguish different degrees of resource slackness and hence may select an item that results in bottleneck conditions for the constraints. Second, the effective gradient measure does not guarantee the feasibility of the selected item and thus additional computations are necessary to exam the solution feasibility in each iteration. Finally, the two existing heuristics for the general MDKP are generalizations based on their respective 0-1 MDKP counterparts and, as such, they inherently increment the value of each decision variable only by one unit in each iteration. This inevitably leads to computational inefficiency when the decision variables can assume large values.

Our allocation heuristic (the order-based allocation rule in the ATO system) is developed on the notion of *effective capacity,* which is intuitively understood as the maximum number of copies of an item that can be accepted if the entire remaining capacity of the knapsack were to be used for that item alone. Our heuristic starts with the feasible solution in which all decision variables are set to zero, and in each iteration selects the item that generates the highest total reward with its effective capacity, and commits a proportion, say α, of its effective capacity to that item. As such, our heuristic *always* generates a feasible solution in each iteration of the algorithm. In essence, our method uses a "max-min" criteria to make variable selection decisions and reflects the multidimensional nature of the general MDKP. In addition, the heuristic depletes the effective capacity at a geometric rate α and in the meantime avoids creating *bottleneck* conditions for other items. The ability of our heuristic to generate a feasible solution and to include multiple copies of the selected item in the solution in each iteration results in superior computational efficiency over other heuristics. We refer to our heuristic as the *primal effective capacity heuristic* (PECH) in the rest of the chapter. It is worth noting that although our heuristic is primarily developed for the component allocation problem in an ATO system, it can be used to solve any general MDKP, as well as the 0-1 MDKP after suitable adjustments.

We first test the effectiveness and efficiency of our heuristic using component allocation problems of ATO systems. Several important managerial observations emerge from our computational results of ATO systems. First, most optimization models in ATO research focus on determining optimal base-stock levels in order to reach a given quality of service, under some simple allocation rules (e.g., FCFS in continuous review models and the fixed priority rule in periodic review models). It is perceived that the customer service level depends solely on the base-stock levels and that the impact of component allocation decisions is marginal. Our findings in this chapter redress this misperception: We show that even when there are abundant on-hand inventories to meet customer orders, the attainable service level may not be realized due to poor component allocation decisions. In other words, ineffective allocation rules can significantly diminish the benefits of risk pooling, which is the underlying philosophy of component commonality in the ATO strategy. Second, a component-based component allocation rule, albeit simple and easy to implement, is generally ineffective unless the service level is very high. Our computational results indicate that, the percentage difference between the optimal reward and the reward collected using a component-based allocation rule can be more than 25%. For higher service levels, this difference is less drastic, but still significant, and can exceed 5%. Indeed, the very nature of the ATO operation, where common components are shared by many different customer orders and each order requires simultaneous availability of several components, implies that the firm should use an order-based allocation rule that focuses on coordinated allocation decisions across different components. Fortunately, our work provides an order-based allocation rule that is both simple and effective and suitable for real-time implementation. Finally, not only does an ineffective allocation rule degrade customer service, it can also lead to high inventory holding costs and thus has the compounded, adverse impact on the firm's profit. This happens if the firm only ships completely assembled orders but still charges holding costs to the components committed to the partially filled orders.

We also test our heuristics on randomly generated general and 0-1 MDKPs. Our computational results further indicate that our heuristic solution method PECH dominates all other heuristics in computational time and provides the best solution among the existing heuristics in 61.2% of the cases for the general MDKP, and in 49% of the cases for the 0-1 MDKP. In particular, PECH is remarkably effective for the problem instances with large numbers of constraints and small-to-medium numbers of decision variables (typically less than 200), usually considered as "hard" multidimensional knapsack problems, and excels when the tightness of the knapsack's constraints improve. This is not surprising, as the "max-min" criterion of our heuristic makes it particularly powerful to handle high-dimensional knapsack problems. Also, it nicely complements the effective gradient methods and its variants, as the "max-sum" criterion used in those methods is generally effective for low-dimensional knapsack problems.

The remainder of this chapter is organized as follows. After a review of the literature in Section 2, we describe the component allocation problem and formulate the associated integer program in Section 3. We propose a component allocation heuristic and its variants in Section 4, and examine their properties in Section 5. Our computational results are reported in Section 6. Finally, we summarize our contributions and discuss the future research in Section 7.

2. Literature Review

2.1 Literature review on ATO systems

Baker et al., 1986, study a simple single-period, two-product ATO system, where each product requires a special component and a common component shared by both products. The objective is to minimize the total safety stock of the components subject to an aggregated type-I service level requirement. They show that the total inventory can be reduced by using the common component, as the result of the risk pooling of component commonality. They also examine an alternative optimization problem, where the objective is again to minimize the total safety stock, but subject to the constraints that each product must satisfy its individual type-I service level requirement. Component allocation decisions are naturally introduced in this problem when the stock of the common component is not enough to meet the demand for both products. The authors derive the analytical expressions for the component stock levels by giving priority to the product with the smaller realized demand. Gerchak et al., 1988, extend the above results to a more general product structure setting and revisits the component allocation problem. Solving a two-stage stochastic program, they derive the optimal rationing policy for the two-product model, which gives priority to the first product with a certain probability. In order to generalize the framework, Gerchak and Henig, 1989, consider the multi-period version of the problem and show that the optimal solution is a myopic policy under certain conditions. Hausman et al., 1998, study a periodic review, multi-component ATO system with independent order-up-to policies where demand follows a multivariate normal distribution. They formulate a nonlinear program aiming at finding the optimal component order-up-to levels, subject to an inventory budget constraint, that maximizes an aggregate type-I service level requirement, here defined as the probability of *joint* demand fulfillment within a common time window. They propose an *equal fractile* heuristic as a solution method to determine the order-up-to levels and test its effectiveness. Schraner, 1995, investigates the capacitated version of the problem and suitably modifies the equal fractile heuristic of Hausman et al., 1998. In addition, Swaminathan and Tayur, 1999, formulated a stochastic program model to study inventory replenishment and component allocation decisions in a multi-period ATO system. In particular, they formulated the component allocation problem as a linear program that can be solved by conventional methods.

There are two papers reported in the literature that directly address the component allocation policies in the ATO system and that are particularly relevant to our research. Zhang, 1997, studies a system similar to that of Hausman et al., 1998; he is interested in determining the order-up-to level of each component that minimizes the total inventory cost, subject to a type-I service level requirement for *each* product. Zhang proposes the *fixed priority* component allocation rule, under which all the product demands requiring a given component are assigned a predetermined priority order, and the available inventory of the component is allocated accordingly. The ranking of products for each component is determined by factors such as product rewards or marketing policies. Agrawal and Cohen, 2000, investigate the *fair shares scheme* as an alternative component allocation policy. The fair shares scheme allocates the available stock of components to product orders, independent of the availability of the other required components. The quantity of the component allocated to a product is determined by the ratio of the realized demand of that product to the total realized demand of all the product orders. They derive the analytical expression for the type-I service level for each product and further determine the optimal component stock levels that minimize the total inventory cost, subject to product service level requirements. It is worth mentioning that both the fixed priority rule and the fair shares scheme are *component-based* allocation rules in the sense that the allocation decisions of a component depend on the product demand of that component alone and are independent of the allocation decisions made for other components. On the one hand, the advantage of a component-based allocation rule is its simplicity: it is easily implemented at the local level and does not require system-wide information. On the other hand, it is conceivable that such a rule could result in sizable "partially" filled orders and long order response times.

As an alternative to these periodic review systems, Song, 1998, studies a continuous-time, base-stock ATO system. She assumes a multivariate Poisson demand process in an uncapacitated system with deterministic lead times. Since orders arrive one by one and the first-come, first-served rule is used to fill customer orders, the component allocation problem is irrelevant. Song expresses the order fill rate, defined as the probability of filling an order instantaneously, by a series of convolutions of one-dimensional Poisson distributions and proposed lower and upper bounds for the order fill rate. Song et al., 1999, use a set of correlated $M/M/1/c$ queues to model the capacitated supply in a continuous-time ATO system. They propose a matrix-geometric solution approach and derive exact expressions for several performance measures. Built upon this model, Xu, 1999, studies the effect of demand correlation on the performance of the ATO system (also see Xu, 2001). Glasserman and Wang, 1998, model capacitated ATO systems using correlated $M^D/G/1$ and $G^D/G/1$ queues. They characterize the trade-offs between delivery lead time and inventory. Wang, 1999, proposes base-stock policies to manage the inventories in these systems at a minimum cost subject to service level requirements. Gallien and Wein, 1998, develop analogous results for a single-product, uncapacitated stochastic assembly system where component replenishment orders are synchronized, and obtained an approximate base-stock policy. Song and Yao, 2000, investigate a problem similar to Gallien and Wein's, but focus on asynchronized systems. Chen et al., 2000, study the configure-to-order (CTO) system, which takes the ATO concept one step further in allowing customers to select a customized set of components that go into the product. They use a lower bound on the fill rate of each product to achieve tractability, and solve the optimization problem by minimizing the maximum stockout probability among all product families subject to an inventory budget. They propose a greedy heuristic as a solution method and test its effectiveness on realistic problem data.

2.2 Literature review on MDKP

The heuristic solution methods for the 0-1 MDKP (MDKP with binary decision variables) have generated a great deal of interest in the literature. Senju and Toyoda, 1968, propose a *dual gradient* method that starts with a possibly infeasible initial solution (all decision variables set to 1) and achieves feasibility by dropping the non-rewarding variables one by one, while following an effective gradient path. Toyoda, 1975, develops a *primal gradient* method that improves the initial feasible solution (all decision variables set to 0) by incrementing the value of the decision variable with the steepest effective gradient. Exploiting the basic idea behind Toyoda's primal gradient method, Loulou and Michaelides, 1979, develop a greedy-like algorithm that expands the initial feasible solution by including the decision variable with the maximum pseudo-utility, a measure of the contribution rate per unit of the aggregate resource consumption of all resources by the decision variable. Incorporating Senju and Toyoda's dual gradient algorithm and Everett's *Generalized Lagrange Multipliers* approach (Everett, 1963), Magazine and Oguz, 1984, propose a heuristic method that moves from the initial infeasible solution towards a feasible solution by following a direction which reduces the aggregate weighted infeasibility among all resource constraints. In addition, Pirkul, 1987, present an efficient algorithm that first constructs a standard 0-1 knapsack problem using the dual variables (known as the *surrogate multipliers*) obtained from the linear programming relaxation of the 0-1 MDKP; he then solves this simpler problem using a greedy algorithm based on the ordering of the return to resource consumption ratios. We refer the reader to the survey paper by Lin, 1998, and the references therein on the results for the 0-1 MDKP and other non-standard knapsack problems.

Unlike the extensive researches that have been conducted for the 0-1 MDKP, the solution approaches for the general MDKP are scarce. To the best of our knowledge, only two heuristics have been reported in the literature that are primarily developed for the general MDKP. Kochenberger et al., 1974, generalize Toyoda's primal gradient algorithm, developed for the 0-1 MDKP, to handle the general MDKP. This method starts with the initial feasible solution and increases one unit of the variable with the largest effective gradient, where the effective gradient of an item is the ratio of the reward of the item to the sum of the portions of slack consumptions of the item over all resource constraints. Pirkul and Narasimhan, 1986, extend the approximate algorithm of Pirkul for the 0-1 MDKP to solve the general MDKP. Their method fixes the variables to their upper bounds in sequential order of their return to consumption ratios until one or more of the constraints of the problem are violated.

3. Problem Description and Formulation

3.1 The system

We consider a periodic review ATO system with m components, indexed by $i = 1, 2, \ldots, m$; and n products, indexed by $j = 1, 2, \ldots, n$. The inventory position of each component is reviewed at the beginning of every review period t, $t = 0, 1, 2, \ldots$. The replenishment of component i is controlled by an *independent* base-stock policy, with the base-stock level for component i denoted by S_i, $i = 1, \ldots, m$. That is, if at the beginning of period t, the inventory position (i.e., inventory on-hand plus inventory on order minus backorders) of component i is less than S_i, then order up to S_i; otherwise, do not order. The independent base-stock policy in general is not optimal in the ATO system, but has been adopted in analysis and in practice due to its simple structure and easy implementation. We assume that the replenishment lead time of component i, denoted by L_i, is a constant integer that can be different for different components. After replenishment orders are received, customer orders for different products arrive. The customer order of product j in period t is denoted by random variable P_{jt}, where $(P_{1t}, P_{2t}, \ldots, P_{nt})$ can be correlated for the same period but are independent and identically distributed (i.i.d.) random vectors across different periods.

Each product is assembled from multiple units of a subset of components. Let b_{ij} be the number of units of component i required for one unit demand of product j, $i = 1, \ldots, m$, $j = 1, \ldots, n$. The system quotes a constant response time window, w_j, for product j and receives a unit reward, r_j, if an order of product j is filled within w_j periods after its arrival, where an order of product j is said to be filled if the order is allocated b_{ij} units of component i, $i = 1, 2, \ldots, m$. Reward r_j is understood as the premium the firm receives by filling a unit of product j demand within its time window. All unfilled orders are completely backlogged.

For a given base-stock policy, the amount of inventory to be allocated to the unfilled demands is determined based on a first-come first-served

(FCFS) order fulfillment discipline. Note that under the FCFS rule, no inventory is committed to the orders received in later periods, unless earlier backlogs for a component are entirely satisfied.

We use the following notation in the rest of the chapter. For $i = 1, \ldots, m$, $j = 1, 2, \ldots, n$ and $t = 0, 1, \ldots$ define,

P_{jt} = Demand for product j in period t;
b_{ij} = Usage rate of component i for one unit demand of product j;
D_{it} = Total demand for component i in period $t = \sum_{j=1}^{n} b_{ij} P_{jt}$;
A_{it} = Replenishment of component i received in period t;
I_{it} = Net inventory (on-hand plus on order minus backlog) of component i at the end of period t;
S_i = Base-stock level of component i;
L_i = Replenishment lead time of component i;
w_j = Response time window of product j;
r_j = Reward rate of filling a unit demand of product j within response time window w_j.

For convenience, we sort the product indices in ascending order of their response time windows so that

$$w_1 \leq w_2 \leq \cdots \leq w_n = w.$$

Next, we derive several identities that will facilitate the formulation of our model. Let $D_i[s, t]$ and $A_i[s, t]$ represent the total demand and total replenishment of component i, $i = 1, 2, \ldots, m$, from period s through period t inclusive. Then

$$D_i[s, t] = \sum_{k=s}^{t} D_{ik} \quad \text{and} \quad A_i[s, t] = \sum_{k=s}^{t} A_{ik}, \qquad \text{for } i = 1, \ldots, m.$$

Based on Hadley and Whitin, 1963, the net inventory of component i at the end of period $t + k$ under the base-stock control S_i, is given by

$$I_{i,t+k} = S_i - D_i[t + k - L_i, t + k], \; i = 1, \ldots, m. \tag{2.1}$$

Since the system uses FCFS to fill orders, using the balance equation, we can relate the ending inventory of component i at periods t and $t + k$ as follows:

$$I_{i,t+k} = I_{it} + A_i[t + 1, t + k] - D_i[t + 1, t + k]. \tag{2.2}$$

Using equations (2.1) and (2.2), we write, for $t + k \geq L_i$,

$$I_{it} + A_i[t + 1, t + k] - D_i[t + 1, t + k] = S_i - D_i[t + k - L_i, t + k]. \tag{2.3}$$

We also know that

$$I_{it} + A_i[t+1, t+k] = I_{i,t-1} + A_i[t, t+k] - D_{it}. \tag{2.4}$$

Substituting (2.4) into (2.3), we reach the following result:

$$I_{i,t-1} + A_i[t, t+k] - D_{it} - D_i[t+1, t+k] = S_i - D_i[t+k-L_i, t+k],$$

which can be further simplified, for $t + k \geq L_i$, $i = 1, \ldots, m$, as

$$\begin{aligned} I_{i,t-1} + A_i[t, t+k] &= S_i - D_i[t+k-L_i, t+k] + D_i[t, t+k] \\ &= S_i - D_i[t+k-L_i, t-1]. \end{aligned}$$

Observe that $I_{i,t-1} + A_i[t, t+k]$ is the net inventory of component i in period $t+k$, after receiving all replenishment orders from periods t to $t+k$, but before allocating any inventory to the orders received after period $t-1$. Due to the FCFS principle in inventory commitment, customer orders received in period t, $P_{1,t}, \ldots, P_{n,t}$, will be filled before the orders received in the subsequent periods. Thus,

$$(S_i - D_i[t+k-L_i, t-1])^+ = \max\{S_i - D_i[t+k-L_i, t-1], 0\}$$

is indeed the on-hand inventory of component i available in period $t + k$ that can be used to fill the orders $P_{1,t}, \ldots, P_{n,t}$, provided that no inventory of component i has been allocated to those orders since their arrival, $k = 0, 1, 2, \ldots, w$.

In steady state, we drop the time index t from our notation and use D_i and P_j to denote the generic versions of D_{it} and P_{jt}. We also represent the stationary version of $D_i[t+k-L_i, t-1]$ by $D_i(L_i - k)$, the total stationary demand of component i in $L_i - k$ periods. In addition, we shall assume that $L_i \geq w$ for all i, where $w = w_n$ is the maximal response time window. This assumption loses no generality since if $L_i < w$ for some i, then the demand for component i from all product orders can be filled before their response time windows and component i can be eliminated from our decisions.

3.2 Integer Programming formulation

We shall formulate the component allocation problem as an integer program. Let

$$\xi = \{(P_1, \ldots, P_n),\ D_i(L_i - k),\ i = 1, \ldots, m,\ k = 0, 1, \ldots, w\}$$

be the collection of random demands, where $P_1, \ldots, P_n$ are the product orders that arrive in the current period (without loss of generality, designate the current period as period 0), and $D_i(L_i - k)$ is the total demand of component i generated in the previous $L_i - k$ periods, $i = 1, 2, \ldots, n$; $k = 0, 1, \ldots, w$. For a demand realization $\boldsymbol{\xi}(\omega) = \{(p_1, \ldots, p_n),\ d_i(L_i - k), i = 1, \ldots, m,\ k = 0, 1, \ldots, w\}$ and given base-stock levels **S**, let $Q_{\mathrm{w}}(\mathbf{S}, \boldsymbol{\xi}(\omega))$ be the maximal total reward attainable from the orders $p_1, p_2, \ldots, p_n$. The performance measure $\beta_{\mathrm{w}}(\mathbf{S})$, which is the long-run average reward ratio, or the percentage of total reward attainable per period, under a base-stock policy given by **S**, can be written as:

$$\beta_{\mathrm{w}}(\mathbf{S}) \quad = \quad 100\% \times \frac{E_{\boldsymbol{\xi}}[Q_{\mathrm{w}}(\mathbf{S}, \boldsymbol{\xi})]}{\sum\limits_{j=1}^{n} r_j E[P_j]}. \tag{2.5}$$

Remember that the aggregated type-II service level of a system is defined as the proportion of all demands satisfied in a period. If the rewards from all products are equal, the associated variables will cancel out in the numerator and denominator of the ratio given in (2.5) and this measure reduces to the aggregated type-II service level. Even if the rewards are not equal, the system might choose to use the type-II service level if it gives equal importance to all products being manufactured. As we mentioned before, the type-II service level provides a better picture of service from customers' perspective (Axsäter, 2000). The following is our integer programming formulation for the component allocation problem:

$$Q_{\mathrm{w}}(\mathbf{S}, \boldsymbol{\xi}(\omega)) \quad = \quad \max\left\{ \sum_{j=1}^{n} \sum_{k=0}^{w_j} r_j x_{jk} \right\} \tag{2.6}$$

subject to

$$\sum_{j=1}^{n} \sum_{\ell=0}^{k} b_{ij} x_{j\ell} \quad \le \quad (S_i - d_i(L_i - k))^+,$$

$$i = 1, 2, \ldots, m \text{ and } k = 0, 1, \ldots, w \tag{2.7}$$

$$\sum_{k=0}^{w} x_{jk} \quad \le \quad p_j, \quad j = 1, 2, \ldots, n \tag{2.8}$$

$$x_{jk} \quad \ge \quad 0 \text{ and integer,}$$

$$j = 1, 2, \ldots, n \text{ and } k = 0, 1, \ldots, w. \tag{2.9}$$

The decision variable x_{jk} is the number of customer orders for product j that are filled k periods after they are received, for $0 \le k \le w$. Observe from (2.6) that we do not collect rewards for the orders filled after their response time windows. The on-hand inventory constraint for component i in (2.7) states that the total allocation of component

i within the first k periods, $0 \le k \le w$, cannot exceed the on-hand inventory of component i, $(S_i - d_i(L_i - k))^+$, $i = 1, 2, \ldots, m$. Let a_i be the shorthand notation for the on-hand inventory of component i. The demand constraint for product j in (2.8) ensures that the total units of product j demand filled within its response time window do not exceed the demand of product j, p_j, $j = 1, 2, \ldots, n$.

Next, we examine the computational complexity of our integer program. Consider a special case where the response time windows of all products are zero, $w_j = 0$, $j = 1, 2, \ldots, n$. For simplicity, let $x_{j0} := x_j$, $j = 0, 1, \ldots, n$. Then the component allocation problem defined in (2.6)-(2.9) is to determine, for given $\mathbf{S}$ and $\boldsymbol{\xi}(\omega)$, the immediate fills $(x_1, \ldots, x_n)$ that maximize the total reward:

$$Q_0(\mathbf{S}, \boldsymbol{\xi}(\omega)) = \max \sum_{j=1}^{n} r_j x_j$$

subject to

$$\begin{aligned} \sum_{j=1}^{n} b_{ij} x_j &\le (S_i - d_i(L_i))^+, \quad i = 1, 2, \ldots, m, \\ x_j &\le p_j, \quad j = 1, 2, \ldots, n, \\ x_j &\ge 0 \text{ and integer}, \quad j = 1, 2, \ldots, n. \end{aligned}$$

This simpler version of our component allocation problem is equivalent to the general MDKP. As we noted earlier, the MDKP is NP-hard (Garey and Johnson, 1979). The reader may easily see the analogies between our allocation problem and the general MDKP, where the product types in the former correspond to the item types in the latter, and the inventory constraints in the former match the knapsack constraints in the latter. More generally, the component allocation problem with positive response time windows is equivalent to a multi-period, multidimensional knapsack problem, under which the items to be placed in the knapsack during the first k periods have to satisfy m capacity constraints for each $k = 0, 1, \ldots, w$.

4. Primal Effective Capacity Heuristics (PECH) for MDKP

In this section, we present our primal effective heuristic to solve various multidimensional knapsack problems. First we consider a general MDKP and provide the PECH$_\alpha$ algorithm as an approximate solution method. We slightly modify this algorithm to handle the component allocation problem stated in (2.6)-(2.9). Then, the PECH algorithm for the 0-1 MDKP follows.

Let $\lfloor q \rfloor$ be the largest integer not exceeding q. Let α, and $0 < \alpha \leq 1$, be a control parameter that determines at what rate the slack of the knapsack is committed to the selected item. The following algorithm, PECH_α, describes an approximate solution procedure for a general MDKP.

ALGORITHM 2.1 (The PECH_α algorithm for solving the general MDKP)

```
BEGIN
   STEP 1  Initialize decision variables $x_j = 0,\ \forall j$;
           Initialize set $E = \{j | x_j = 0,\ \forall j\}$;
           Initialize capacities of resources $\bar{a}_i = a_i,\ \forall i$;
           Initialize upper bounds of decision variables $\bar{p}_j = p_j,\ \forall j$;
   STEP 2  Compute effective capacity for item $j$:
              $\bar{y}_j = \min_i \left\{ \left\lfloor \frac{\bar{a}_i}{b_{ij}} \right\rfloor : b_{ij} > 0 \right\},\ \forall j \in E.$
           If $\bar{y}_j = 0, \forall j \in E$, then go to END, otherwise go to STEP 3.
   STEP 3  Compute $r_j \times \bar{y}_j,\ \forall j \in E$ and select $j^* = \arg\max_{j \in E}\{r_j \times \bar{y}_j\}$.
   STEP 4  Compute the increment of item $j^*$:
              $y_{j^*} = \min\{\bar{p}_{j^*}, \max\{1, \lfloor \alpha \bar{y}_{j^*} \rfloor\}\}$.
   STEP 5  Update the values of decision variables: $x_{j^*} \leftarrow x_{j^*} + y_{j^*}$;
           Update remaining capacities of constraints:
              $\bar{a}_i \leftarrow \bar{a}_i - b_{ij^*} \times y_{j^*},\ \forall i$;
           Update slacks of decision variables: $\bar{p}_{j^*} \leftarrow \bar{p}_{j^*} - y_{j^*}$;
           Update set $E$: If $\bar{p}_{j^*} = 0$ or $\alpha = 1$, set $E \leftarrow E - \{j^*\}$;
           If $E = \emptyset$, go to END; otherwise go to STEP 2.
END
```

A brief explanation of PECH_α is in order: STEP 1 sets the initial feasible solution to zero and also initializes resource constraints and upper bounds of decision variables to their respective initial values. STEP 2 computes, for each item $j \in E$, its *effective capacity* $\bar{y}_j$, which is the maximum number of copies of item j that can be accepted with the available capacity of the knapsack. STEP 3 computes $r_j \times \bar{y}_j$, the maximum reward of item j were the entire remaining capacity of the knapsack dedicated to the item, $\forall j$, and selects the item j^* that has the largest attainable reward. STEP 4 accepts either $\max\{1, \lfloor \alpha \bar{y}_{j^*} \rfloor\}$ units of item j^*, which is $\alpha \times 100\%$ of the effective capacity of item j^*, or its current upper bound $\bar{p}_{j^*}$, whichever is smaller. Finally, STEP 5 updates the information and returns to STEP 2 for the next iteration. Note that the algorithm satisfies the integrality conditions for the decision variables and always generates a feasible solution.

The greediness of PECH_α is determined by the selection of the *greedy coefficient* α. As α increases, the rate at which the available resources are consumed by the selected item in each iteration increases. In the extreme case, $\alpha = 1$, the selected item can use its maximal effective capacity and thus maximum feasible number of copies will be included to the solution in a single iteration. For small α, the algorithm becomes conservative in its resource commitment to the selected item. The algorithm might even accept only a single copy of an item in an iteration as the resource capacities become tight. The use of the greedy coefficient reduces the possibility of creating bottleneck conditions for certain constraints, which may result in reduced rewards in future iterations. Moreover, the heuristic is likely to select different items in consecutive iterations since the maximum rewards may change over time. It is evident that PECH$_\alpha$ is a computationally more efficient algorithm over other general MDKP algorithms, since it depletes the effective capacity geometrically at rate α and does not need to examine the feasibility of the solution after each step.

Next, we modify Algorithm 2.1 to solve the component allocation problem with positive time windows. In the algorithm, we use the component allocation problem terminology for clarity.

ALGORITHM 2.2 (The PECH$_\alpha^w$ algorithm for the component allocation problem with given **S**, $\boldsymbol{\xi}(\omega)$ and response time windows, $w_1 \le w_2 \cdots \le w_n = w$.)

```
BEGIN
    STEP 0  **Initialize** time index $k = 0$;
            **Set** the unfilled orders $p_j$, $\forall j$.
    STEP 1  **Initialize** the set of unfilled orders in period $k$:
                $E_k = \{j \mid w_j \ge k \text{ and } p_j > 0,\ \forall j\}$;
            **Set** filled orders in period $k$: $x_{jk} = 0$, $\forall j \in E_k$.
            **Initialize** the on hand inventory in period $k$:
                $a_i = \left(S_i - d_i(L_i - k)\right)^+ - \sum_{\ell=0}^{k} \sum_{j=1}^{n} b_{ij} x_{j\ell}$, $\forall i$.
    STEP 2  **Compute** the effective capacity of product $j$ in period $k$:
                $\bar{y}_j = \min_i \left\{ \left\lfloor \frac{A_i}{b_{ij}} \right\rfloor : b_{ij} > 0 \right\}$ $\forall j \in E_k$.
    STEP 3  If $\bar{y}_j = 0$ $\forall j \in E_k$ **go to** STEP 7; otherwise **go to** STEP 4.
    STEP 4  **Compute** $r_j \times \bar{y}_j$ $\forall j \in E_k$ and select $j^* = \arg\max_{j \in E_k} \{r_j \times \bar{y}_j\}$.
            **Break ties** by selecting the product with the largest number of
            unfilled orders.
    STEP 5  **Fill** $y_{j^*} = \min\left\{p_{j^*}, \max\{1, \lfloor \alpha \bar{y}_{j^*} \rfloor\}\right\}$ units of product $j^*$.
    STEP 6  **Update** the on hand inventory $A_i \leftarrow A_i - b_{ij^*} y_{j^*}$, $\forall i$with $b_{ij^*} > 0$;
            **Update** the unfilled orders: $p_{j^*} \leftarrow p_{j^*} - y_{j^*}$;
            **Update** the filled orders in period $k$: $x_{j^*k} \leftarrow x_{j^*k} + y_{j^*}$;
            **Update** the unfilled order types: If $p_{j^*} = 0$ or $\alpha = 1$,
                **set** $E_k \leftarrow E_k - \{j^*\}$;
```

```
        If $E_k = \emptyset$, go to STEP 7; otherwise go to STEP 2.
STEP 7  Increment $k \leftarrow k + 1$. If $k > w$, go to END; otherwise go to
            STEP 1.
END
```

In this algorithm, the on hand inventory has to be updated for each given period k, $0 \leq k \leq w$, to account for the replenishment. The steps taken for each k are similar to those of Algorithm 2.1, and we repeat the procedure for $k = 0, 1, \ldots, w$. After the response time window of a product expires, the algorithm no longer allocates inventory to the unfilled orders of that product, if any. However, those orders will be filled prior to the orders received subsequently, due to FCFS principle.

Next, we show how Algorithm 2.1 can be used to solve the 0-1 MDKP. Our algorithm for the 0-1 MDKP deviates from PECH_α for the general MDKP in two aspects. First, the greedy coefficient a used in Algorithm 2.1 is no longer needed, as decision variables can only take binary values. Second, the increment of the selected item is always one unit and once an item is included in the solution it is immediately removed from *E,* the set of unselected items. As a result, PECH is terminated in at most n iterations.

ALGORITHM 2.3 (The PECH algorithm for solving the 0-1 MDKP)

```
BEGIN
  STEP 1  Initialize decision variables $x_j = 0,\ \forall j$;
          Initialize the set of unselected items $E = \{j | x_j = 0,\ \forall j\}$;
          Initialize resource capacities $\bar{a}_i = a_i,\ \forall i$;
  STEP 2  Compute effective capacity for item $j$:
              $\bar{y}_j = \min_i \left\{ \left\lfloor \frac{\bar{a}_i}{b_{ij}} \right\rfloor : b_{ij} > 0 \right\},\ \forall j \in E.$
          If $\bar{y}_j = 0, \forall j \in E$, then go to END, otherwise go to STEP 4.
  STEP 3  Compute $r_j \times \bar{y}_j,\ \forall j \in E$ and select $j^* = \arg\max_{j \in E} \{r_j \times \bar{y}_j\}$.
  STEP 4  Let $x_{j^*} \leftarrow 1$;
          Update the remaining resource constraints $\bar{a}_i \leftarrow \bar{a}_i - b_{ij^*},\ \forall i$;
          Update the set of unselected items $E \leftarrow E - \{j^*\}$;
          If $E = \emptyset$, go to END; otherwise go to STEP 2.
END
```

5. Properties of the PECH Heuristic

We first examine the properties of PECH_α and identify the conditions under which it locates the optimal solution. We then consider the computational complexity of PECH_α.

The next proposition states that under certain *agreeable conditions,* PECH_α^w in Algorithm 2.2 shares the same property as the optimal allocation policy (proof of this proposition is given in Akçay and Xu, 2002).

PROPOSITION 2.4 *If* $r_{j_1} \geq r_{j_2}$, $w_{j_1} \leq w_{j_2}$ *and* $b_{i,j_1} \leq b_{i,j_2}$ *for* $i = 1, 2, \ldots, m$, *then for any realization of demand,* $\xi(\omega) = \{p_1, \ldots, p_n, d_i(L_i - k), i = 1, \ldots, m, k = 1, \ldots, w)$, *the following is true:*

1 *Let* $\pi^* = \{x^*_{jk},\ j = 1, 2, \ldots, n, k = 0, \ldots, w\}$ *be the optimal solution, where* x^*_{jk} *is the number of customer orders of product* j *that are filled* k *periods after their arrival. Then,*

$$\sum_{k=1}^{w_{j_1}} x^*_{j_1,k} < p_{j_1} \Longrightarrow \sum_{k=1}^{w_{j_1}} x^*_{j_2,k} = 0.$$

2 *Let* $\pi = \{x_{jk},\ j = 1, 2, \ldots, n, k = 0, \ldots, w\}$ *be the solution of the* $PECH^w_\alpha$ *heuristic. Then,*

$$\sum_{k=1}^{w_{j_1}} x_{j_1,k} < p_{j_1} \Longrightarrow \sum_{k=1}^{w_{j_1}} x_{j_2,k} = 0.$$

In the special case, $r_j = 1$, $w_j = w$, and $b_{ij} \in \{0, 1\}$ for all i and j, Proposition 2.4 implies that both the optimal policy and the PECH^w_α heuristic satisfy the *small-order-first* property. That is, if product-*L* requests a single unit of each component in *L* and if $L \subseteq K \subseteq \{1, 2, \ldots, m\}$, then both policies will not fill the orders of product *K* unless the orders of product-*L* are fully satisfied.

A policy is said to be an *index policy* if in each period it fills customer orders according to a predetermined sequence of product indices, as long as their response time windows have not expired. The next result follows directly from Proposition 2.4.

COROLLARY 2.5 *Suppose the following* agreeable *conditions are satisfied:*

$$\begin{aligned} &r_1 \geq r_2 \geq \ldots \geq r_n, \\ &w_1 \leq w_2 \leq \ldots \leq w_n, \\ &b_{i1} \leq b_{i2} \leq \ldots \leq b_{in} \qquad \textit{for } i = 1, 2, \ldots, m. \end{aligned}$$

Then,

1 *The index policy* $\{1, 2, \ldots, n\}$, *which in each period fills customer orders in the increasing order of the product indices, starting with the product of the smallest index whose response time window has not expired, is optimal.*

2 *The* $PECH^w_\alpha$ *heuristic becomes the index policy* $\{1, 2, \ldots, n\}$ *and hence is optimal.*

Clearly, the the optimal allocation policy does not fill the demand of a product after its response time window expires. Examining Algorithm 2.2, it is seen that PECH_α^w shares the same property.

We now turn our attention to the computational complexity of the algorithms we discussed in the previous section. We show that PECH_α for the general MDKP is a polynomial-time algorithm with the computational complexity $O(mn^2)$ if $\alpha = 1$ and is a *pseudo polynomial-time* algorithm if $0 < \alpha < 1$. We also show that PECH for the 0-1 MDKP is a polynomial-time algorithm with the computational complexity $O(mn^2)$.

We first consider PECH_α, with $\alpha = 1$, for the general MDKP described in Algorithm 2.1. The effect of the initialization step, STEP 1, is negligible on complexity. STEPS 2 and 3 of the algorithm require $O(mn)$ basic operations. For $\alpha = 1$, the maximum feasible number of copies of the selected item will be included to the solution in a single iteration. Thus, the algorithm will be repeated once for each of the n items in the worst case. This implies that PECH_1 for the general MDKP is a polynomial-time algorithm with the computational complexity $O(mn^2)$.

We next consider PECH_α, with $0 < \alpha < 1$, for the general MDKP. Similar to the case $\alpha = 1$, each iteration requires $O(mn)$ basic operations. Let k_j be the number of iterations in which item j is selected by PECH_α, $j = 1, 2, \ldots, n$, in STEP 4. Clearly, k_j depends on the values of the problem parameters and the greedy coefficient α and its exact value is difficult to determine. However, we recognize that $k_j \leq min\{p_j, K_j\}$, where K_j is the number of times that PECH_α would have selected item j if the entire knapsack capacity were to be used for the item alone and $p_j = \infty$, $j = 1, 2, \ldots, n$. We shall derive an expression for K_j and use it to bound k_j.

Let $y_j = \min_i\{\lfloor \frac{a_i}{b_{ij}} \rfloor\}$ be the maximal number of copies of item j, $j = 1, 2, \ldots, n$, that can be packed in the knapsack, given initial capacities a_i, $i = 1, \ldots, m$. Let

$$R = \{j : \alpha y_j < 2\}.$$

Clearly, if $j \in R$, then PECH_α will accept only one unit of item j at each iteration. We thus have

$$k_j \leq \min\{p_j, K_j\} = \min\{p_j, y_j\}, \quad j \in R. \tag{2.10}$$

Now consider $j \notin R$. Since in each iteration, the heuristic commits $\alpha \times 100\%$ of its remaining capacity to the selected item, the number of units accepted in the kth selection of item j is $\alpha(1-\alpha)^{k-1}y_j$ (for notation simplicity we ignore the integrality constraint), if the entire effective capacity were used for item j. Therefore, the number of iterations in which *multiple* copies of item j are accepted is given by

$$K_j^M = \max\{k : \alpha(1-\alpha)^{k-1}y_j \geq 2\} = \frac{\log(\frac{2}{\alpha y_j})}{\log(1-\alpha)} + 1, \ \ j \notin R. \quad (2.11)$$

Now let K_j^S be the number of iterations in which PECH$_\alpha$ accepts a single copy of item j. It is clear that K_j^S must equal to the remaining effective capacity of item j after the K_j^Mth selection, $(1-\alpha)^{K_j^M} y_j$. We thus have:

$$K_j^S = (1-\alpha)^{K_j^M} y_j \leq \frac{2}{\alpha}, \qquad j \notin R, \qquad (2.12)$$

where the last inequality is the result of the definition of K_j^M, which implies $\alpha(1-\alpha)^{K_j^M} y_j < 2$, $j \notin R$. Combining (2.11) and (2.12), we obtain, for $j = 1, 2, \ldots, n$,

$$\begin{aligned} k_j &\leq \min\{p_j, K_j\} = \min\{p_j, K_j^M + K_j^S\} \\ &\leq \min\{p_j, \frac{\log(\frac{2}{\alpha y_j})}{\log(1-\alpha)} + \frac{2}{\alpha} + 1\}, \qquad j \notin R. \qquad (2.13) \end{aligned}$$

From (2.10) and (2.13), we obtain the following bound for $\sum_j k_j$, the total number of iterations under PECH$_\alpha$:

$$\begin{aligned} \sum_{j=1}^{n} k_j &= \text{Total number of iterations of the heuristic} \\ &\leq \sum_{j \in R} \min\{p_j, y_j\} + \sum_{j \notin R} \min\left\{p_j, \frac{\log(\frac{2}{\alpha y_j})}{\log(1-\alpha)} + \frac{2}{\alpha} + 1\right\}. \end{aligned}$$

Furthermore, if we consider the worst case when the algorithm accepts one unit of each item in every iteration ($R = \{1, 2, \ldots, n\}$), the following bound is valid from (2.10):

$$\sum_{j=1}^{n} k_j \leq \sum_{j=1}^{n} \min\{p_j, y_j\}.$$

In summary, we have:

PROPOSITION 2.6

1 *If* $\alpha = 1$, *then PECH*$_\alpha$ *for the general MDKP, given in Algorithm 2.1, is a* polynomial-time *algorithm with the computational complexity of* $O(mn^2)$.

2 *If* $0 < \alpha < 1$, *then* $PECH_\alpha$ *for the general MDKP, given in Algorithm 2.1, is a* pseudo polynomial-time *algorithm with its running time bounded by the computational complexity of*

$$O\left(mn \sum_{j \notin R} \min\{p_j, y_j\} + mn \sum_{j \in R} \min\{p_j, \frac{\log(\frac{2}{\alpha y_j})}{\log(1-\alpha)} + \frac{2}{\alpha} + 1\}\right)$$

and

$$O\left(mn \sum_{j=1}^{n} \min\{p_j, y_j\}\right).$$

The heuristic of Kochenberger et al., 1974, has a computational complexity of $O(mn^2 \sum_{j=1}^{n} \min\{up_j, y_j\})$. Therefore, even in the worst case, our heuristic is computationally more efficient than Kochenberger et al.'s. On the other hand, the heuristic of Pirkul and Narasimhan, 1986, is possibly an exponential-time algorithm, if a simplex algorithm is used to solve the LP relaxation of the problem.

We can generalize Proposition 2.6 to the PECH_α^w heuristic given in Algorithm 2.2. For simplicity, we only state the computational complexity result for the PECH_1^w heuristic.

PROPOSITION 2.7 *The* $PECH_1^w$ *heuristic with response time windows* $w_1 \leq \cdots \leq w_n = w$, *stated in Algorithm 2.2, is a polynomial-time algorithm with the computational complexity* $O(mn^2 w)$.

Now we consider PECH for the 0-1 MDKP, described in Algorithm 2.3. Since the decision variables in the 0-1 MDKP are binary variables, STEP 2 to STEP 3 of the algorithm, which require $O(mn)$ basic operations, are repeated once for each of the n items in the worst case. Hence we state:

PROPOSITION 2.8 *The algorithm of PECH, given in Algorithm 2.3, is a polynomial-time algorithm with the computational complexity of* $O(mn^2)$.

The heuristics of Toyoda, 1975, Loulou and Michaelides, 1979, and Magazine and Oguz, 1984, have a complexity of $O(mn^2)$, which is the same as ours. However, the heuristic of Pirkul, 1987, has the potential of an exponential complexity of $O(m2^n)$, if a simplex algorithm is used to solve the LP relaxation of the problem.

6. Computational Results

In this section, we present our computational results that compare the performance of our heuristics against other existing approximate solution methods. We coded the algorithms in C language, incorporating CPLEX 7.5 optimization subroutines, and performed the computations on a dual processor WinNT server, with two Intel Pentium III Xeon 1.0 GHz processors and 2GB of RAM.

In our results, we report the following performance statistics:

1 *Mean:* This is the mean value of the percentage error, where the percentage error of each heuristic is measured as

$$\text{percentage error} = 100\% \times \frac{Z^* - Z_{\text{heuristic}}}{Z^*},$$

where Z^* and $Z_{\text{heuristic}}$ are the optimal and heuristic solutions of the problem, respectively.

2 *CPU:* This is the percentage of the mean CPU time each heuristic takes to obtain the solution compared with the average CPU time to obtain the optimal solution.

6.1 Computational results for ATO problems

In our first set of computations, we test the performance of our heuristic in solving the component allocation problem in an ATO system. We use the general MDKP algorithms of Kochenberger et al., 1974, and Pirkul and Narasimhan, 1986, and component allocation rules of Zhang, 1997, and Agrawal and Cohen, 2000, as benchmark solutions.

We first consider a PC assembly system with realistic problem data from the IBM Personal Systems Group Chen et al., 2000. In this system, a family of six desktop computers are to be assembled from a set of seventeen components. The demand for each product is normally distributed with a mean of 100 and a standard deviation of 20. A base-stock policy is used to manage the replenishment of component inventories, and the base-stock level of each component is set such that it covers the mean component demand over its effective lead time, plus a certain level of safety stock. Therefore, for component i,

$$S_i = (L_i + 1)E[D_i] + z[(L_i + 1)Var(D_i)]^{1/2}. \qquad (2.14)$$

The safety stock of the component can be adjusted for the desired service level by fine tuning the safety factor z. As the value of z increases, the system is better protected against demand variation, hence the service level also increases. The length of the response time window is zero for all products. Other problem parameters are given in Table 2.1.

For each test case, we generate 1000 random demand instances for each product. Our computational results with different levels of z are

Table 2.1. Problem parameters

			Products					
		j	1	2	3	4	5	6
		r_j	1,363	1,595	1,765	1,494	1,494	1,628
i	Components	L_i	*Component Usage Rates*					
1	Shell	5	1	1	1	1	1	1
2	Shell (Common Parts 1)	8	1	1	1	1	1	1
3	Shell (Common Parts 2)	8	1	1	1	1	1	1
4	Processor 450 MHz	12	1	0	0	0	0	0
5	Processor 500 MHz	12	0	1	0	0	0	0
6	Processor 550 MHz	12	0	0	1	1	1	0
7	Processor 600 MHz	12	0	0	0	0	0	1
8	Memory 64MB	15	1	1	1	1	1	1
9	Hard drive 6.8GB	18	1	1	0	0	0	1
10	Hard drive 13.5GB	18	0	0	1	1	1	0
11	Hard drive (Common Parts)	8	1	1	1	1	1	1
12	Software Pre-load 1	4	1	1	1	0	1	0
13	Software Pre-load 2	4	0	0	0	1	0	0
14	CD-ROM	10	0	0	1	0	0	0
15	CD-ROM (Common Parts)	10	0	0	1	0	0	0
16	Video Graphics Card	6	1	1	1	1	1	0
17	Ethernet Card	10	0	0	1	0	0	0

provided in Table 2.2. Our findings clearly indicate that $\mathrm{PECH}_{0.1}$ is the best heuristic in terms of the mean error among all heuristics. It provides solutions within 0.72% of the optimal in only 0.18% of the time required by the optimal solution, on average over all the randomly generated instances for this problem. The CPU time of PECH_α increases as the value of α decreases, as expected. Even when $\alpha = 0.1$, its CPU time is drastically better than that of KOCH and PIR_G. Again, the batch allocation feature of our heuristic is the primary source of this difference. Moreover, our results reveal the ineffectiveness of component-based allocation rules in ATO systems. Under tight capacity constraints (smaller values of z), the fair shares and fixed priority allocation rules cannot adequately address the component allocation decisions, and have unacceptable differences with the optimal. Even for moderately higher service levels, the gap between the optimal and their respective solutions might reach 5%.

Table 2.2. Performance statistics for the IBM PC assembly system

	Component Allocation Rules									
	$\mathrm{PECH}_{1.0}$		$\mathrm{PECH}_{0.5}$		$\mathrm{PECH}_{0.1}$		FS		FP	
z	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
0.0	4.41	0.02	2.66	0.10	1.46	0.23	16.53	0.04	31.79	0.06
0.5	3.67	0.01	2.38	0.06	1.42	0.26	9.42	0.04	17.35	0.04
1.0	2.05	0.03	1.56	0.07	0.95	0.20	4.74	0.03	7.35	0.03
1.5	1.03	0.02	0.71	0.07	0.41	0.18	1.65	0.03	2.13	0.04
2.0	0.23	0.02	0.22	0.06	0.10	0.13	0.33	0.03	0.34	0.04
2.5	0.04	0.03	0.03	0.05	0.02	0.10	0.05	0.02	0.04	0.04

	General MDKP Heuristics			
	KOCH		PIR_G	
z	Mean	CPU	Mean	CPU
0.0	8.23	0.63	2.42	1.24
0.5	6.81	0.85	2.47	1.12
1.0	4.79	0.92	1.50	1.08
1.5	2.45	1.01	0.44	1.05
2.0	0.72	1.16	0.13	0.89
2.5	0.15	1.18	0.02	0.82

PECH_α	Our primal effective capacity heuristic
FS	Fair shares rule of Agrawal and Cohen, 2000
FP	Fixed priority rule of Zhang, 1997
KOCH	MDKP heuristic of Kochenberger et al., 1974
PIR_G	MDKP heuristic of Pirkul and Narasimhan, 1986

Next, we consider component allocation problem instances in randomly generated ATO systems. We assume that the response time windows of all products are equal to zero. The leadtime of component i is constructed using a discrete uniform distribution between 1 and 10, $i = 1, \ldots, 20$. The reward rate for each product is uniformly distributed between 10 and 20, and the product demands are normally distributed with the parameters given in Table 2.3.

Table 2.3. Parameters of the product demand distribution

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
Mean	100	40	150	80	200	20	120	50	250	60
Std. Dev.	30	10	25	20	50	5	30	10	50	15

The usage rate of component i for product j is generated from a discrete distribution with the probability mass function $\pi_{hj} = P\{b_{ij} = h\}$, $h = 0, 1, \ldots, 5$ and $j = 1, \ldots, 10$, as follows:

$$\Pi = \begin{bmatrix} \frac{3}{4} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{3} & \frac{1}{2} & \frac{1}{4} & \frac{1}{2} & \frac{1}{6} & \frac{1}{4} & \frac{1}{8} & \frac{1}{4} & \frac{1}{10} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{6} & \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{10} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{10} \end{bmatrix}$$

We define the *size* of a product as the total number of components required to assemble one unit of the product. Notice that, with the above distribution, the size of the product with a larger index is *expected* to rise. The base-stock level of component i is set using equation (2.14). We generate 100 problem instances and report the performance statistics. The test results indicate that $\text{PECH}_{0.1}$ provide the best solution among all the heuristic methods for every problem instance and that it is very close to the optimum. The solution quality of PECH_1 is still acceptable, but not as good as that of KOCH and PIR_G, which provide excellent solutions. On the other hand, the component-based allocation rules once again suffer under tight capacity levels.

6.2 Computational results for general MDKP

In this section, we randomly generate multiple resource allocation problems, which can be modelled as general MDKP. We select the number of resource constraints, the number of decision variables, and the

Table 2.4. Performance statistics for the randomly generated component allocation problems

	Component Allocation Rules									
	PECH$_{1.0}$		PECH$_{0.5}$		PECH$_{0.1}$		FS		FP	
z	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
0.0	2.32	0.02	0.55	0.06	0.58	0.29	11.38	0.04	23.02	0.04
1.0	1.39	0.03	0.16	0.10	0.14	0.41	3.63	0.04	5.71	0.03
2.0	0.62	0.06	0.05	0.16	0.03	0.59	0.56	0.02	0.81	0.04

	General MDKP Heuristics			
	KOCH		PIR$_G$	
z	Mean	CPU	Mean	CPU
0.0	0.84	1.04	0.61	1.24
1.0	0.44	1.45	0.28	1.18
2.0	0.24	2.46	0.09	0.76

slackness of the knapsack constraints as the design parameters in our numerical tests. By varying the values of these parameters, we are able to generate a wide range of random problems. More specifically, we generate these design parameters and other problem data as follows:

- The number of resource constraints m and the number of decision variables n: We choose four levels for m, $m = 10, 50, 100$ and 200, and five levels for n, $n = 10, 50, 100, 200$ and 400;

- The greedy coefficient α: We set $\alpha = 1, 0.5$ and 0.1. This allows us to observe how the solution quality and computational time of PECH$_\alpha$ are affected by the greediness of the algorithm;

- We randomly generate the values of b_{ij}, r_j and p_j from discrete uniform distributions, $b_{ij} \sim \mathit{Unif}[0,9]$, $r_j \sim \mathit{Unif}[1,100]$ and $p_j \sim \mathit{Unif}[1,200]$, for all j.

- The *slackness ratio* σ_i: To generate the values of a_i with various levels of resource capacities, we adopt the *slackness ratio* σ_i introduced by Zanakis, 1977:

$$\sigma_i = a_i / \sum_{j=1}^{n} b_{ij} p_j \quad \text{for } i = 1, 2, \ldots, m \tag{2.15}$$

  Note that resource capacities increase as the slackness ratio σ_i increases. In our computations, we first generate b_{ij} and p_j as described above, and substitute them in equation (2.15) with σ_i to determine a_i. We use i.i.d. uniform random distributions to obtain the value of σ_i for each resource constraint. We study four different cases: $\sigma_i \sim Unif(0.4, 0.6), \sigma_i \sim Unif(0.6, 0.8), \sigma_i \sim Unif(0.8, 0.1.0)$ and $\sigma_i \sim Unif(0.4, 1.0)$;

This setup for the computational experiments gives us 4 × 5 × 4 = 80 test cases. For each test case, we generate 100 instances of b_{ij}, r_j, p_j and σ_i from their respective distributions. Results from all 80 cases can be found in Akçay et al., 2002. In Tables 2.5, 2.6 and 2.7 we present the performance statistics of the heuristics with varying problem parameters, namely the number of resource constraints, the number of decision variables and the slackness ratios of resource constraints. We construct these tables by simply taking the averages of our results with respect to the fixed m, n or σ_i, over the values of other parameters.

We note from Table 2.5 that the performance of all heuristics deteriorates as the number of resource constraints increases, both in terms of *Mean* and *CPU.* However, this trend is more striking for $\mathrm{PIR_G}$. This is to be expected, as it is particularly designed to solve the general MDKP with a small number of constraints Pirkul and Narasimhan, 1986. Note also that KOCH provides the best solution for the problems with a small number of constraints.

Table 2.6 shows that the solution quality of the heuristics increases as the number of decision variables increases, but the CPU time also increases. Observe that, in terms of the mean error, $\mathrm{PECH}_{0.5}$ outperforms all other heuristics for the instances with $n = 10, 50, 100$, but underperforms KOCH and PIR_G for the instances with $n = 200$, 400. A plausible explanation of this result is that the batch assignment feature of PECH can exhaust the knapsack's capacity prematurely when the number of decision variables are large. We also notice that the performance of KOCH is relatively robust, whereas that of PIR_G is very sensitive to the number of decision variables. In terms of the CPU time, $\mathrm{PECH}_{0.5}$ dominates both KOCH and PIR_G, particularly for the problem with a large number of decision variables.

As seen in Table 2.7, when the expected value of the slackness of the constraints increases, with the variance fixed (compare the first three rows of Table 2.7), the solution quality of all heuristics improves. Notice that the performances of both PECH and KOCH are relatively robust against the tightness of slackness ratios, but PIR_G is not. It is particularly striking to observe that the CPU time of our heuristics decreases as the expected value of the slackness of the constraints increases, whereas those of KOCH and $\mathrm{PIR_G}$ worsen. We believe that this phenomenon can be explained by the unique capability of PECH to deplete its remaining capacity at a geometric rate α. When the knapsack has ample slacks, the batch assignment feature of PECH shows its power, as the other two heuristics can only add a single copy of the selected item to the solution in each iteration. Finally, when the variance of the slackness of the constraints increases, with the mean fixed (compare the second and the fourth rows of Table 2.7), both performance measures suffer in all heuristics.

In Table 2.8, we display the heuristics with the minimum mean error for various problem sizes in terms of the number of constraints and number of decision variables, averaged over a wide range of slackness ratios. As noted earlier, PECH dominates the other heuristics when the number of variables is relatively smaller compared to the number of constraints. Therefore, it is desirable to use our heuristic for general knapsack problems with high dimensions.

Table 2.5. General MDKP – Performance statistics for different number of resource constraints

	PECH$_{1.0}$		PECH$_{0.5}$		PECH$_{0.1}$		KOCH		PIR$_G$	
m	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
10	1.44	0.18	0.78	0.30	0.88	0.49	0.56	0.56	1.64	0.54
50	2.58	0.68	1.18	1.14	1.32	1.93	1.34	2.28	4.28	2.53
100	2.94	1.24	1.25	2.02	1.36	3.63	1.79	4.23	5.70	4.82
200	3.70	2.83	1.31	4.23	1.39	7.11	2.39	8.50	6.93	10.18

Table 2.6. General MDKP – Performance statistics for different number of decision variables

	PECH$_{1.0}$		PECH$_{0.5}$		PECH$_{0.1}$		KOCH		PIR$_G$	
n	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
10	8.20	0.06	1.67	0.09	1.74	0.17	3.65	0.16	17.25	0.05
50	1.77	0.16	1.08	0.25	1.21	0.42	1.40	0.44	3.05	0.28
100	1.34	0.40	1.04	0.64	1.19	1.04	1.11	1.25	1.66	1.07
200	1.10	1.39	0.97	1.97	1.08	3.42	0.84	4.04	0.86	4.26
400	0.92	4.15	0.87	6.65	0.97	11.41	0.60	13.56	0.38	16.93

Table 2.7. General MDKP – Performance statistics for different slackness ratios

	PECH$_{1.0}$		PECH$_{0.5}$		PECH$_{0.1}$		KOCH		PIR$_G$	
S$_i$	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
(0.4,0.6)	4.24	1.82	1.84	2.86	1.98	4.63	2.27	2.33	7.08	2.63
(0.6,0.8)	2.21	0.94	0.83	1.72	0.89	2.82	1.28	3.21	4.32	3.79
(0.8,1.0)	1.03	0.72	0.33	1.11	0.34	1.95	0.64	5.54	2.23	6.72
(0.4,1.0)	3.18	1.44	1.51	2.00	1.73	3.76	1.88	4.49	4.92	4.94

Table 2.8. General MDKP – Best heuristic (in terms of Mean) for different number of constraints and variables

	n				
m	10	50	100	200	400
10	PECH$_{0.5}$	KOCH	KOCH	PIR$_G$	PIR$_G$
50	PECH$_{0.5}$	PECH$_{0.5}$	PECH$_{0.5}$	PIR$_G$	PIR$_G$
100	PECH$_{0.5}$	PECH$_{0.5}$	PECH$_{0.5}$	PIR$_G$	PIR$_G$
200	PECH$_{0.5}$	PECH$_{0.5}$	PECH$_{0.5}$	PECH$_{0.5}$	PIR$_G$

6.3 Computational results for 0-1 MDKP

We choose the system configuration similarly as in Section 6.2 for the general MDKP, except that we let $p_j \equiv 1$ for all j. We compare the performance of our heuristic against the 0-1 MDKP heuristics of Senju and Toyoda, 1968; Toyoda, 1975; Loulou and Michaelides, 1979; Magazine and Oguz, 1984; and Pirkul, 1987. Results for all 80 cases are given in Akçay et al., 2002. Table 2.9 shows that PIR provides the minimal mean error when m is small, whereas PECH becomes dominant when m increases. On the other hand, Table 2.10 indicates that PECH provides the minimal mean error when n is small, but gradually yields its leading role to PIR (when $n \geq 100$), SEN (when $n \geq 200$) and MAG (when $n = 400$) as n increase. It is worth noting that even though the performance of PECH as compared with that of several others deteriorates for large n, it still provides excellent approximate solutions. For example, when $n = 400$, the mean errors of the best heuristic, PIR, and PECH are 0.32% and 0.78%, respectively. Table 2.11 shows that PECH outperforms all other heuristics both in terms of solution quality and CPU against all levels of the slackness ratio. The robustness of PECH is particularly striking since it is primarily designed to solve the general MDKP.

Table 2.12 presents the heuristics with the minimum mean error for various problem sizes in terms of the number of constraints and number of decision variables. Clearly, PECH outperforms the other heuristics when the number of constraints are relatively larger than the number of variables, which can be observed in the lower triangle region of the table.

Table 2.9. 0-1 MDKP – Performance statistics for different number of resource constraints

	$m = 10$		$m = 50$		$m = 100$		$m = 200$	
	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
PECH	0.92	0.12	1.23	0.45	1.43	0.86	4.28	1.70
SEN	1.50	0.12	2.29	0.47	2.60	0.91	6.21	1.79
TOY	2.02	0.13	2.53	0.52	2.84	0.99	6.47	1.98
LOU	2.42	0.23	2.51	1.03	2.73	2.03	6.27	4.03
MAG	2.36	0.15	4.08	0.64	5.30	1.25	6.25	2.48
PIR	0.86	0.39	1.84	1.83	3.04	3.61	7.27	7.19

SEN: Heuristic of Senju and Toyoda, 1968
TOY: Heuristic of Toyoda, 1975
LOU: Heuristic of Loulou and Michaelides, 1979
MAG: Heuristic of Magazine and Oguz, 1984
PIR: Heuristic of Pirkul, 1987

Table 2.10. 0-1 MDKP – Performance statistics for different number of decision variables

	$n = 10$		$n = 50$		$n = 100$		$n = 200$		$n = 400$	
	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
PECH	5.49	0.04	1.56	0.10	1.11	0.25	0.89	0.81	0.78	2.71
SEN	10.39	0.04	2.48	0.10	1.46	0.26	0.86	0.84	0.57	2.87
TOY	8.03	0.04	2.83	0.10	2.31	0.29	2.13	0.94	2.02	3.16
LOU	8.55	0.04	3.08	0.18	2.27	0.57	1.86	1.88	1.65	6.48
MAG	16.02	0.03	3.05	0.13	1.72	0.35	1.04	1.16	0.67	3.98
PIR	12.17	0.04	2.15	0.22	1.07	0.79	0.55	3.06	0.32	12.16

7. Conclusions

We summarize our results in this chapter as follows.

1 We model the component allocation problem in an ATO system as an integer program. Unlike most studies reported in the lit-

Table 2.11. 0-1 MDKP – Performance statistics for different slackness ratios

	$S_i \in (0.4, 0.6)$		$S_i \in (0.6, 0.8)$		$S_i \in (0.8, 1.0)$		$S_i \in (0.4, 1.0)$	
	Mean	CPU	Mean	CPU	Mean	CPU	Mean	CPU
PECH	3.28	1.10	1.48	0.67	0.62	0.47	2.48	0.89
SEN	4.88	0.48	2.58	0.71	1.32	1.16	3.83	0.93
TOY	5.38	0.53	2.67	0.78	1.15	1.28	4.66	1.03
LOU	5.44	1.06	2.70	1.57	1.17	2.60	4.62	2.09
MAG	7.13	0.66	3.81	0.97	1.98	1.60	5.07	1.29
PIR	5.34	1.87	2.86	2.79	1.44	4.64	3.38	3.71

Table 2.12. 0-1 MDKP – Best heuristic (in terms of Mean) for different number of constraints and variables

			n		
m	10	50	100	200	400
10	PECH	PIR	PIR	PIR	PIR
50	PECH	PECH	PIR	PIR	PIR
100	PECH	PECH	PECH	PIR	PIR
200	PECH	PECH	PECH	PECH	PIR

erature, we adopt the the long-run average reward ratio, which reduces to the type-II service level for identical reward rates, as the performance measure. We believe that, for the ATO system with batch product demands, this performance measure reflects the true meaning of service from customers' perspective.

2 The component allocation problem in an ATO environment has not been well studied in the literature. The two reported component allocation rules are component-based and are not effective in the ATO environment. Here, we formulate the component allocation problem as a general MDKP and propose the first order-based heuristic and show that it can be solved in either polynomial or pseudo-polynomial time. Intensive testing indicates that our heuristic is robust, efficient and effective and is suitable for real-time implementation.

3 Our results provide several insights that enhance understanding of effective management in ATO systems. First, ineffective allocation rules can significantly diminish the benefits of risk pooling, the underlying philosophy of component commonality in the ATO strategy. Second, the coordinated, order-based component alloca-

tion rule significantly outperforms the component-based allocation rules. Finally, an ineffective allocation rule not only degrades customer service, but it can also lead to high inventory holding costs and thus result in a compounded, adverse impact on system performance.

4 As reported by Lin, 1998, the studies toward the solution approaches to general MDKP are scarce and most solution procedures are efficient only when the number of resource constraints is small. To the best of our knowledge, the only approximate solution methods for general MDKP are the effective gradient method by Kochenberger et al., 1974, and the surrogate relaxation method by Pirkul and Narasimhan, 1986. Lin also states that for future researches, "the development of polynomial time algorithms which generate the near-optimal solution through heuristic approaches remains attractive". Based on a comprehensive computational study, we demonstrate that the new heuristic proposed in this chapter significantly improves computational efficiency of the existing general and 0-1 MDKP heuristics and generates robust and near-optimal solutions, especially for high dimensional knapsack problems with small-to-moderate numbers of decision variables.

References

Agrawal, N. and Cohen, M.A. (2000). Optimal material control in an assembly system with component commonality. Working paper, Department of Decision and Information Sciences, Santa Clara University, Santa Clara, CA 95053.

Akçay, Y., Li, H., and Xu, S.H. (2002). An approximate algorithm for the general multidimensional knapsack problem. Working paper, Pennsylvania State University.

Akçay, Y. and Xu, S.H. (2002). Joint inventory replenishment and component allocation optimization in an assemble-to-order system. Working paper, Pennsylvania State University.

Axsäter, S. (2000). *Inventory Control.* Kluwer Academic Publishers, Boston/Dordrecht/London.

Baker, K.R., Magazine, M.J., and Nuttle, H.L.W. (1986). The effect of commonality on safety stock in a simple inventory model. *Management Science,* 32(8):982–988.

Chen, F., Ettl, M., Lin, G., and Yao, D.D. (2000). Inventory-service optimization in configure-to-order systems: From machine-type models to building blocks. Research Report RC 21781 (98077), IBM Research Division, IBM T.J. Watson Research Center, York Heights, NY 10598.

Everett, H. (1963). Generalized langrange multiplier method for solving problems of optimum allocation of resources. *Operations Research,* 2:399–417.

Gallien, J. and Wein, L.M. (1998). A simple and effective procurement policy for stochastic assembly systems. Working paper, MIT Sloan School.

Garey, M.R. and Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman, San Francisco.

Gerchak, Y. and Henig, M. (1989). Component commonality in assemble-to-order systems: Models and properties. *Naval Research Logistics,* 36:61–68.

Gerchak, Y., Magazine, M.J., and Gamble, A.B. (1988). Component commonality with service level requirements. *Management Science,* 34(6):753–760.

Glasserman, P. and Wang, Y. (1998). Leadtime inventory trade-offs in assemble-to-order systems. *Operations Research,* 46:858–871.

Hadley, G. and Whitin, T.M. (1963). *Analysis of Inventory Systems.* Prentice Hall, Englewood Cliffs, NJ.

Hausman, W.H., Lee, H.L., and Zhang, A.X. (1998). Joint demand fulfillment probability in a multi-item inventory system with independent order-up-to policies. *European Journal of Operational Research,* 109:646–659.

Kochenberger, G.A., McCarl, B.A., and Wyman, F.P. (1974). A heuristic for general integer programming. *Decision Sciences,* 5:36–44.

Lee, H.L. and Tang, C.S. (1997). Modelling the costs and benefits of delayed product differentiation. *Management Science,* 43:40–53.

Lin, E.Y. (1998). A bibliographical survey on some well-known nonstandard knapsack problems. *INFOR,* 36(4):274–317.

Loulou, R. and Michaelides, E. (1979). New greedy-like heuristics for the multidimensional 0-1 knapsack problem. *Operations Research,* 27:1101–1114.

Magazine, M.J. and Oguz, O. (1984). A heuristic algorithm for the multidimensional zero-one knapsack problem. *European Journal of Operational Research,* 16:319–326.

Pirkul, H. (1987). A heuristic solution procedure for the multiconstraint zero-one knapsack problem. *Naval Research Logistics,* 34:161–172.

Pirkul, H. and Narasimhan, S. (1986). Efficient algorithms for the multiconstraint general knapsack problem. *IIE Transactions,* pages 195–203.

Schraner, E. (1995). Capacity/inventory trade-offs in assemble-to-order systems. Working paper, Department of Operations Research, Stanford University, Stanford, CA 94306.

Senju, S. and Toyoda, Y. (1968). An approach to linear programming with 0-1 variables. *Management Science,* 15(4):B196–B207.

Song, J. (1998). On the order fill rate in a multi-item, base-stock inventory system. *Operations Research,* 46(6):831–845.

Song, J., Xu, S.H., and Liu, B. (1999). Order fulfillment performance measures in an assemble-to-order system with stochastic leadtimes. *Operations Research,* 47:131–149.

Song, J. and Yao, D.D. (2000). Performance analysis and optimization of assemble-to-order systems with random leadtimes. preprint.

Swaminathan, J.M. and Tayur, S.R. (1999). *Stochastic Programming Models for Managing Product Variety.* in Quantitative Models for Supply Chain Management, Tayur, Ganashan and Magazine (eds), Kluwer Academic Publishers, 585-624, Boston/Dordrecht/London.

Toyoda, Y. (1975). A simplified algorithm for obtaining approximate solutions to zero-one programming problems. *Management Science,* 21(12):1417–1427.

Wang, Y. (1999). Near-optimal base-stock policies in assemble-to-order systems under service level requirements. Working paper, MIT Sloan School.

Xu, S.H. (1999). Structural analysis of a queueing system with multi-classes of correlated arrivals and blocking. *Operations Research,* 47:264 –276.

Xu, S.H. (2001). *Dependence Analysis of Assemble-to-Order Systems.* in Supply Chain Structures: Condition, Information and Optimization, Song and Yao (eds), Kluwer Academic Publishers, 359-414, Boston/Dordrecht/London.

Zanakis, S.H. (1977). Heuristic 0-1 linear programming: Comparisons of three methods. *Management Science,* 24:91–103.

Zhang, A.X. (1997). Demand fulfillment rates in an assemble-to-order system with multiple products and dependent demands. *Production and Operations Management,* 6(3):309–323.

Chapter 3

IMPROVING SUPPLY CHAIN PERFORMANCE THROUGH BUYER COLLABORATION

Paul M. Griffin, Pınar Keskinocak, and Seçil Savaşaneril
School of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive
Atlanta, Georgia 30332-0205
{pgriffin.pinar,secil}@isye.gatech.edu

Abstract We study alternative buyer strategies in markets where procurement costs are effected by economies of scale in the suppliers' production costs and by economies of scope in transportation. We consider buyer strategies with different levels of collaboration, namely, (i) no collaboration among buyers or buyer divisions, (ii) internal collaboration among the purchasing organizations of the same buyer enabled by an internal intermediary, and (iii) full collaboration among multiple buyers enabled by a third party intermediary. We test these buyer strategies under different market conditions and provide insights on when intermediaries have a significant impact on the economic value as well as the buyer surplus.

1. Introduction

In recent years, we have seen a rapid growth of electronic markets (e-markets or exchanges) which bring buyers and suppliers together and enable many-to-many transactions. Depending on who owns and operates them, we can divide e-markets into two categories (Elmaghraby and Keskinocak (2002)): (1) Private e-markets that are owned by a company (commonly referred to as a private trading exchange (PTX)) or a group of companies (co-op or consortia exchanges (CTX)) who are either a buyer(s) or seller(s) in the market, and (2) independent trading exchanges (ITX) that are owned and operated by an independent entity (i.e., not wholly/partially owned by a buyer or seller in the market). Examples of such e-markets are GE Plastics (geplastics.com), which is a PTX, Transplace (transplace.com), which is a consortium exchange formed by the merger of logistics business units of the six largest publicly-held truckload carriers, and Chemconnect (chemconnect.com), an ITX specializing in chemicals and plastics. The benefits of e-markets compared to traditional channels include reduced transaction costs and access to a larger base of potential buyers and suppliers.

Although e-markets began with great fanfare, they have not, at least so far, lived up to their advanced billing. This is especially true for ITXs. Among the reasons why ITXs have failed are privacy concerns, and more importantly, the sellers' concerns about being compared solely on price. PTXs, on the other hand, have gained popularity, as more companies strive to streamline their interactions with their supply chain partners. PTXs have the advantage of giving more control to the owner company and enabling information-sharing and other types of collaboration among participants. In order for ITXs and CTXs to be successful, they should offer more than decreased transaction costs or one-stop shopping convenience. All types of exchanges would benefit significantly from collaboration and decision-support tools for the trading process (Stackpole (2001)), (Keskinocak and Tayur (2001)).

In many companies, purchasing is done by multiple functional divisions (or purchasing organizations within the company) which either act independently or have minimal interaction with each other. For example, until very recently purchasing was done locally by managers at each of the Dial Corp.'s sites. It was typical for buyers at one facility to buy the same raw material as a buyer at another Dial plant from two different suppliers at different prices. This approach was ineffective in taking advantage of Dial's volume and corporate-wide buying power (Reilly (2002)). Similarly, until 1997, purchasing at Siemens Medical Systems was done locally, where buyers at Siemens' ultrasound, electromedical, computer tomography, magnetic resonance imaging and angiography divisions independently bought the components and material that their individual plants needed and rarely communicated with each other. There was no pooling of component demand for leveraging purposes (Carbone (2001)). Recently both companies have moved with great success towards centralized procurement, which allows collaboration among internal purchasing units.

In addition to internal collaboration within a company, there is also a growing interest in collaboration among different companies on various supply chain functions, such as demand forecasting, product design, transportation, and procurement. Such inter-company collaboration is sometimes enabled by market intermediaries in the industry. For example, the Internet-based logistics network, Nistevo, consolidates orders from multiple shippers and creates round-trip or dedicated tours. This reduces the costs of the carriers due to better utilization of truck capacity and in turn results in lower prices for the shippers (Strozniak (2001)). For example, Land O'Lakes Inc. has saved $40,000 a month by coordinating its shipping routes with companies such as Georgia-Pacific (Keenan and Ante (2002)).

Motivated by the current practices and the potential benefits of collaborative procurement, we study the performance of buyer strategies with different levels of collaboration under various market conditions. We assume buyers have multiple functional divisions responsible for purchasing. To analyze the benefits of collaboration, we study procurement strategies under three collaboration models: (i) *No collaboration:* Buyer divisions and suppliers trade through traditional sales channels, via one-to-one transactions. No information flow or collaboration exists among the functional divisions of a buyer or among multiple buyers. (ii) *Internal collaboration:* Functional divisions of a buyer collaborate internally. (iii) *Full collaboration:* A third party intermediary enables collaboration among different buyers and allows the participants to achieve benefits from both economies of scale and scope due to reduced fixed production and transportation costs. We refer to the total set of trades among the buyers and suppliers as a *matching*. The three models are shown in Figure 3.1. Note that the connectivity requirements (and hence, related transaction or search costs) between the buyers and the suppliers decrease as the collaboration level increases.

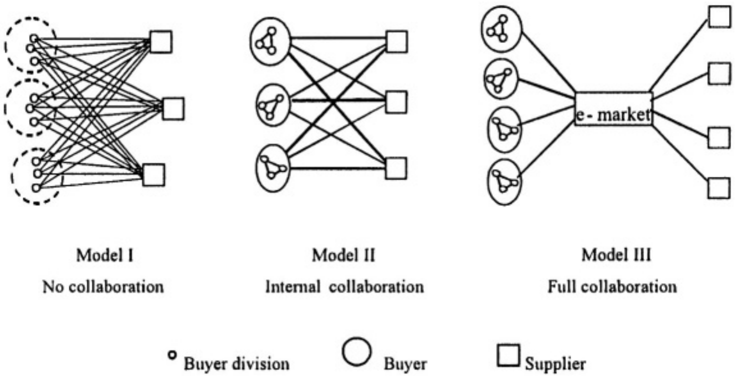


Figure 3.1. Three models of collaboration among buyers and buyer divisions

We conduct numerical experiments to analyze the effectiveness of buyer strategies under different market conditions. We characterize the markets based on three main attributes: the capacity level compared to the total demand in the market, the sellers' relative fixed and variable production costs, and the relative fixed and variable transportation costs. We compare the strategies based on the following performance measures: the percentage of satisfied demand, total surplus of the buyers, average surplus per unit, and the average cost per unit. Buyer surplus can be regarded as some form of buyer profit, which we will formally define in Section 3.

This chapter is organized as follows. In Section 2 we provide a literature review. Alternative buyer strategies with different collaboration levels are presented in Section 3. Experimental results and managerial insights are given in Section 4 and Section 5 followed by a concluding discussion in Section 6.

2. Literature Review

Interactions among participants of a supply chain can be analyzed along several directions. One line of related research focuses on decentralized versus shared information. When the information is decentralized, studies are primarily on constructing different mechanisms to enable coordination in a two-stage setting and to eliminate inefficiencies stemming from double marginalization. Cachon and Zipkin (1999) and Lee et al. (2000) analyze coordination mechanisms in the form of rebates or transfer payments. Weng (1995) considers a system where coordination is established through quantity discounts and franchise fees. Jin and Wu (2001) study supply chain coordination via transfer payments in the presence of e-market intermediaries. See Cachon (2003) for an analysis of coordinating contracts under different supply chain settings.

Another line of research focuses on collaboration. Internet and technology have made information sharing possible at every stage, and this leads to different collaborative efforts in supply chains. Examples include vendor managed inventory (VMI), just-in-time distribution (JITD), and collaborative planning, forecasting and replenishment (CPFR), in which trading partners such as vendors and retailers collaborate vertically. In VMI systems the vendor is given autonomy in replenishing the retailer's orders and in turn manages the retailer's inventory with fewer stockouts and at low levels. In a pilot effort of CPFR, Wal-Mart and vendors Lucent and Sara Lee shared event and point of sales information to jointly forecast sales in 1997 (`http://www.cpfr.org`). There is a growing interest in supply chain literature on analyzing the benefits of VMI in supply chains, for example see Cetinkaya and Lee (2000), and Cheung and Lee (2002). Relatively little research exists on collaboration through forecasting. Aviv (2001) studies the benefits of collaborative forecasting with respect to decentralized forecasting.

Horizontal collaboration differs from vertical in the sense that it considers collaboration among only those on the buyer or supplier side. Horizontal coordination and collaboration in the supply chain enabled by quantity discounts is studied by Gurnani (2001) in a single supplier two buyer setting. Some existing research considers the interaction of buyers and suppliers from a resource allocation perspective. Ledyard, Banks and Porter (1989) test allocation mechanisms with uncertain resources and indivisible demand. The results indicate that high efficiency could be obtained if collaboration is enabled among buyers. To the best of our knowledge there has not been much research on the loss of efficiency due to lack of horizontal collaboration. In this chapter, we study the interaction between multiple buyers and multiple suppliers, where horizontal collaboration is enabled by a central mechanism (or intermediary).

A similar problem is studied by Kalagnanam, Davenport and Lee (2001), where the motivation came from electronic markets in the paper and steel industries. The authors consider an e-market in which buyers and suppliers submit bid and ask prices for multiple units of a single product. They show that the problem of determining the clearing price and quantity under different assignment constraints can be solved in polynomial time when demand is divisible, but is NP-hard when demand is not divisible.

We extend the work of Kalagnanam, Davenport and Lee (2001) in several directions. We consider multiple products, rather than a single product, where each supplier needs to decide how to allocate its limited capacity among these multiple products. Furthermore, we consider fixed costs of production and transportation which lead to economies of scale and scope. Finally, in addition to the "centralized" e-market (where all the buyers and suppliers are available in the market at the same time and the buyer-supplier assignments are done centrally) we study two other scenarios where buyers arrive to the market sequentially and select suppliers on a first come first served basis.

3. Modeling Procurement Decisions under Various Degrees of Collaboration

We model a market with multiple buyers and suppliers where multi-unit transactions for multiple items take place. Buyers have multiple functional divisions where each buyer division is responsible for the procurement of different items. These functional divisions may or may not collaborate in the procurement process to pool their purchasing power.

To model buyers' behavior in the market, we assume that buyers or buyer divisions arrive with requests for quotes (RFQ) for each item they want to buy. We assume that buyers initiate the trades by submitting RFQs to the suppliers. A buyer requests that her entire demand for an item is satisfied from a single supplier. Hence, a supplier would respond to a buyer's RFQ only if he has enough production capacity to satisfy the buyer's entire demand for that item. Such all-or-nothing buyer behavior is observed in several industries for various reasons. Splitting an order across multiple suppliers complicates order tracking and transportation arrangements. In addition, order splitting might lead to inconsistency in quality. For example, in the paper industry the quality of the paper produced by different machines is slightly different, which may create problems in printing. Similarly, in carpet manufacturing, carpet produced at different times or locations has slight color variations which can be noticed when the carpet is installed.

Each buyer has a reservation price for each item, which is the maximum price for the purchase of all the units demanded by that buyer (not per unit of the item). A buyer's *surplus* for an item is defined as the buyer's reservation price for that item minus the final contracting price for the entire demand of the buyer for that item. With the goal of maximizing her surplus, if there are no quotes (bids) that are acceptable to a buyer at a given time, she may leave the market and come back later with the hope of getting a better quote.

By evaluating the quotes offered by the suppliers (if any) buyers decide which supplier to choose for each item. A supplier might produce multiple types of items and has limited production capacity to be shared among these items. A supplier's cost for an item consists of four components:

- The manufacturing setup cost for that item (fixed production cost). A supplier initiates production and incurs a setup cost for an item only if a buyer places an order for that item.
- Variable production cost per unit.
- Fixed cost of transportation.
- Variable transportation cost per unit.

In responding to buyer RFQs, suppliers use a cost plus pricing scheme, i.e., set prices to cover the fixed and variable costs and leave enough *profit* *margin* for profits. Despite its limitations, cost-plus pricing is commonly used in various industries. For example, in the logistics industry, 33% of third-party logistics companies (3PL) in North America used cost-plus pricing in 2000 (Smyrlis (2000)).

To keep the exposition simple, we ignore the *profit margin* component and focus only on the cost component, i.e., the *bid price* quoted by a supplier is obtained by adding up the cost components. However, as we will explain in the following sections, what the buyer pays in the end *(contract price)* might be lower than what the supplier quotes originally.

We assume that the buyers select suppliers based on price alone. Although price is an important criterion in supplier selection, most buyers also consider other factors, such as quality and delivery time reliability, while selecting a supplier. However, we focus on commodity procurement, where multiple vendors with similar quality and delivery performance exist.

In the remainder of this chapter we assume that each buyer division is responsible for the procurement of one item; hence, the index for items and buyer divisions is the same, see Table 3.1. We also assume that the different divisions of the same buyer are located in the same region, which implies that the locations of the divisions are close enough to allow for consolidation of orders for transportation and thus for price discounts from the carriers; e.g., the "region" can be a state, or the south-east region of the United States. The pricing structure enables buyers to obtain economies of scale and scope. As more buyers place orders with the same supplier for the same item, the associated fixed production cost for each buyer decreases (economies of scale). As a single buyer places orders at the same supplier for multiple items, the associated fixed transportation cost per unit decreases (economies of scope). This type of cost (or price) structure can also be interpreted as a volume discount.

The demand quantities of the buyers and initial capacities available at the suppliers represent the total demand and total supply in the market. Initially there is no production setup at the suppliers. As buyers accept supplier bids and make contracts for the items, suppliers initiate production.

3.1 No Collaboration

In this market structure, we model traditional marketplaces, where neither the functional divisions of a buyer nor different buyers in the market collaborate with each other. As discussed earlier, many firms have uncoordinated purchasing divisions. For example, until recently

Table 3.1. Glossary of Notation

i:	index for buyers, $i \in I$
j:	index for items (or buyer divisions), $j \in J$
k:	index for suppliers, $k \in K$
Div_{ij}:	buyer i, division j
d_{ij}:	demand of buyer i for item j
D_j:	total demand in the market for item j, $\sum_i d_{ij}$
q_{ijk}:	quantity of item j produced at supplier k upon receiving an RFQ from buyer i
t_{jk}:	total quantity of item j produced at supplier k
res_{ij}:	reservation price of buyer i for the total demanded quantity of item j
fpc_{jk}:	fixed production cost for item j at supplier k
vpc_{jk} :	variable production cost per unit for item j at supplier k
ftc_{ik} :	fixed cost for transportation between buyer i and supplier k (might also include the fixed transaction costs)
vtc_{ik} :	variable cost for transportation per unit between buyer i and supplier k
c_{jk}:	capacity required to produce one unit of item j at supplier k
tc_k:	total capacity at supplier k
cap_k:	available capacity at supplier k, upon receiving an RFQ

Chevron's procurement structure was fragmented and decentralized in which "people at many different locations were buying many materials, (often the same materials) from their favorite suppliers, or on an as-needed basis" (Reilly (2001)).

We assume that the functional divisions arrive sequentially and independently to the market. Therefore the marketplace can be thought of as a queue of buyer divisions each with an RFQ for a specific item. When a buyer division makes a contracting decision, she is unaware of the other buyer divisions' demands or procurement decisions (including the ones both from the same and different companies), or about the current order status at the suppliers. After submitting an RFQ, a buyer division makes the contracting decision only based on the unit prices quoted by the suppliers. Once the contracting decision is made and the purchase order is submitted by a buyer division, another buyer division arrives to the market and submits an RFQ.

Buyer divisions' decisions in contracting depend on the kind of information they receive from the suppliers. A supplier can provide either a *pessimistic* or an *optimistic* quote to a buyer. When providing a pessimistic quote, the supplier regards that buyer as if she will be the last one to contract for that item. When providing an optimistic quote (OPT), the supplier assumes that he will supply all the demand in the market for that item. The optimistic quote is a lower bound on the final contract price, whereas the pessimistic quote is an upper bound.

Given an RFQ by buyer division j of company i for d_{ij} units of item j, the supplier computes the pessimistic quote (bid price per unit) as follows:

$$bid_{ijk} = \frac{fpc_{jk}}{q_{ijk} + d_{ij}} + vpc_{jk} + \frac{ftc_{ik}}{\sum_{m \in S} d_{im}} + vtc_{ik} \qquad (3.1)$$

The first two terms in equation (3.1) correspond to the unit production cost, $P_k(d_{ij})$. The fixed production cost for an item is shared among multiple buyer divisions (from different companies) who placed orders for that item with the supplier. q_{ijk} is the total quantity for item j already contracted at supplier k upon receiving the RFQ of company i division j.

The last two terms of equation (3.1) correspond to the unit transportation cost, $T_k(d_{ij})$. The fixed transportation cost is shared among multiple buyer divisions from the same company who placed orders (for different items) with the same supplier. The set S contains buyer division j and the other buyer divisions of company i that have already contracted with supplier k.

Note that if a buyer division is the first one to place an order for an item at a supplier, then she is quoted all the fpc. Similarly, when a supplier receives an RFQ for the first time from a buyer division of a particular company, he incorporates all the ftc in the bid.

To compute the optimistic quote, supplier k needs to first compute (an upper bound on) the maximum total quantity of orders for item j he could produce upon receiving an RFQ, which is:

$$Q_{ijk} = \min\left\{D_j, q_{ijk} + \frac{cap_k}{c_{jk}}\right\}$$

The maximum total quantity is bounded by the minimum of two terms. The first is the total demand for item j in the market. The second is the quantity of item j already produced by supplier k plus the maximum additional quantity of item j that can be produced by supplier k.

Therefore the following term is a lower bound on the fixed production cost per unit:

$$\max\left\{\frac{fpc_{jk}}{D_j}, \frac{fpc_{jk}}{q_{ijk} + \frac{cap_k}{c_{jk}}}\right\}.$$

Upon receiving an RFQ, supplier k computes the following optimistic quote for buyer i per unit of item j:

$$\text{OPT}_{ijk} = \frac{fpc_{jk}}{Q_{ijk}} + \frac{ftc_{ik}}{\sum_{m \in S} d_{im}} + vpc_{jk} + vtc_{ik} \tag{3.2}$$

In the remainder of the chapter, we assume that the supplier provides a pessimistic quote unless otherwise stated.

If buyer i contracts with supplier k, the final contract price she pays per unit of item j is:

$$price_{ijk} = \frac{fpc_{jk}}{t_{jk}} + vpc_{jk} + \frac{ftc_{ik}}{\sum_{m \in S} d_{im}} + vtc_{ik}, \tag{3.3}$$

where t_{jk} is the total quantity of item j produced at supplier k in the final matching. The contract price of an item has a similar structure to the bid price for that item. As the number of buyers contracting with a supplier for the same item increases, the unit price to be paid by a buyer decreases. Therefore, in the end the price paid by a buyer might be lower than the quoted price.

The practice of lower final prices paid by the buyers compared to the initial bids offered by the suppliers is commonly observed in group purchasing programs. For example, the price of a product goes down as more buyers join the group and agree to buy that product at the current posted price. Although some buyers may have joined the group (and committed to purchasing) while the price was higher, in the end all the buyers pay the final, lowest price.

In the final matching, the surplus of buyer i contracted with supplier k for item j is:

$$surplus_{ij} = (res_{ij} - price_{ijk} \cdot d_{ij}) \tag{3.4}$$

Note that when considering the supplier bids the buyer division multiplies the bid price with the total demand for the item to evaluate her surplus.

3.1.1 Buyer Strategies.

A buyer division submits RFQs to the suppliers for the item she demands and chooses a supplier with the goal of maximizing her surplus. We consider the following buyer strategies for accepting or rejecting a bid.

1 If some of the supplier bids are lower than her reservation price, the buyer division accepts the minimum bid.

2 If all the bids are higher than the buyer division's reservation price,

(a) she accepts the minimum bid with probability α.

(b) she rejects all the bids with probability 1-α, and

i with probability β, she leaves the market permanently.

ii with probability $1-\beta$ she returns to the market (i.e., joins the end of the queue) since there is a possibility that the minimum bid the buyer receives later is lower than the current minimum bid. The buyer stays in the market until her surplus becomes positive or the bid prices stop decreasing, whichever happens first.

Having defined a general scheme, we consider the following six buyer strategies resulting from specific choices of α and β. In each of these, a buyer division makes a contract with the supplier that offers the minimum bid, if that bid is below her reservation price. Otherwise, the buyer division:

Accept the minimum bid, MIN ($\alpha = 1$) Contracts with the lowest bid supplier.

Myopic Strategy, MYOPIC ($\alpha = 0$, $\beta = 1$) Leaves the market.

Leave and possibly return later, LOQ ($\alpha = 0, 0 < \beta < 1$) Leaves the market with probability $0 < \beta < 1$, returns to the market and joins the end of the queue with probability $1 - \beta$.

Leave and return later, Q ($\alpha = 0$, $\beta = 0$) Returns to the market and joins the end of the queue.

Accept the lowest bid or leave and return later, AOQ ($0 < \alpha < 1$, $\beta = 0$) Accepts the minimum bid with probability α. With probability 1-α she rejects all the bids and joins the end of the queue.

Minimum optimistic bid, MOB Accepts the minimum optimistic bid, $\min_k\{\text{OPT}_{ijk}\}$.

Example: Consider a marketplace where there are four buyer divisions, Div_{ij}, $i = 1, 2$, $j = 1, 2$ (two companies with two divisions each), three suppliers, S_k, $k = 1, \ldots, 3$ and two items, I_1 and I_2. A buyer division j is responsible for item I_j, $j = 1, 2$. The buyer divisions arrive to the market in the following order: (1) company 1 division 1, (2) company 2 division 1, (3) company 1 division 2, (4) company 2 division 2. Buyer divisions use the *accept the minimum bid* strategy for contracting decisions. For simplicity we assume that the suppliers are uncapacitated. The information regarding buyer divisions (Div), suppliers (S) and items (I) is listed in the Tables 3.2-3.8.

Table 3.2. Demand and reservation prices

	Div_{11}	Div_{12}	Div_{21}	Div_{22}
Demand for I_1	40	-	10	-
Demand for I_2	-	25	-	75
Res. price for I_1	2000	-	500	-
Res. price for I_2	-	3750	-	7500

Company 1 division 1 places an RFQ for item 1. The quoted bids per unit of item 1 by supplier 1, supplier 2 and supplier 3 are 18.75, 131.37 and 19.00 respectively (refer to equation (3.1)). Supplier 1 wins the contract for item 1.

Next, company 2 division 1 arrives to the market and submits an RFQ for item 1. Supplier 1 has already initiated production for item 1. Therefore supplier 1 reflects only some portion of the fixed production cost in the quote, whereas supplier 2 and supplier 3 reflect all the fixed production cost in their quotes. The quotes per unit of item 1 by suppliers 1, 2 and 3 are 24.00, 486.00 and 30.00 respectively. Company 2 division 1 contracts with supplier 1 for item 1.

Table 3.3. Fixed and variable production costs

	S_1	S_2	S_3
I_1	$fpc_{11} = 100$ $vpc_{11} = 10$	$fpc_{12} = 140$ $vpc_{12} = 10$	$fpc_{13} = 100$ $vpc_{13} = 10$
I_2	$fpc_{21} = 9125$ $vpc_{21} = 10$	$fpc_{22} = 100$ $vpc_{22} = 10$	$fpc_{23} = 4605$ $vpc_{23} = 10$

Table 3.4. Fixed and variable transportation costs

	S_1	S_2	S_3
B_1	$ftc_{11} = 50$ $vtc_{11} = 5$	$ftc_{12} = 4515$ $vtc_{12} = 5$	$ftc_{13} = 60$ $vtc_{13} = 5$
B_2	$ftc_{21} = 50$ $vtc_{21} = 7$	$ftc_{22} = 4570$ $vtc_{22} = 5$	$ftc_{23} = 50$ $vtc_{23} = 5$

Next, company 1 division 2 places an RFQ for item 2. While quoting the bids, supplier 1 incorporates only some portion of the fixed transportation cost, since supplier 1 has already contracted with division 1 of the same company for item 1. The bids quoted by suppliers 1, 2 and 3 are 380.77, 199.60, 201.60, respectively. For item 2, company 1 division 2 contracts with supplier 2 and is charged for two different fixed transportation costs by both supplier 1 and supplier 2. The surplus she obtains for item 2 is $res_{12} - bid_{122} \cdot d_{12} = 3750 - 199.60 \cdot 25 = -1240.00$. Although the surplus value is negative, since the strategy under consideration is *accept the minimum bid,* this does not impose any restriction on contracting.

Finally company 2, division 2 submits an RFQ for item 2. Supplier 2 has initiated production for item 2. The quotes for item 2 by suppliers 1, 2 and 3 are 139.25, 76.93, 77.07, respectively. For item 2, company 2 division 2 contracts with supplier 2. She is also charged for two different fixed transportation costs. Bid prices and contracted suppliers are shown in Table 3.5.

Table 3.5. Bid Prices

	S_1	S_2	S_3
Div_{11}	$\mathbf{bid_{111} = 18.75}$	$bid_{112} = 131.75$	$bid_{113} = 19.00$
Div_{12}	$bid_{121} = 380.77$	$\mathbf{bid_{122} = 199.60}$	$bid_{123} = 201.60$
Div_{21}	$\mathbf{bid_{211} = 24.00}$	$bid_{212} = 486.00$	$bid_{213} = 30.00$
Div_{22}	$bid_{221} = 139.25$	$\mathbf{bid_{222} = 76.93}$	$bid_{223} = 77.07$

The contract prices paid by the buyers are shown in Table 3.6 and the total surplus is 2115.25. The matching for the no collaboration example is shown in Figure 3.2(a). In the following sections we analyze the same example under different collaboration models.

Table 3.6. Contract prices and Surplus

	$price_{ijk}$	$surplus_{ij}$	$surplus_i$
Div_{11}	$\frac{100}{40+10} + 10 + \frac{50}{40} + 5 = 18.25$	$2000 - 18.25 \cdot 40 = 1270.00$	105.00
Div_{12}	$\frac{100}{25+75} + 10 + \frac{4515}{25} + 5 = 196.60$	$3750 - 196.60 \cdot 25 = -1165.00$	
Div_{21}	$\frac{100}{10+40} + 10 + \frac{50}{10} + 5 = 22.00$	$500 - 22.00 \cdot 10 = 280.00$	2010.25
Div_{22}	$\frac{100}{25+75} + 10 + \frac{4570}{75} + 5 = 76.93$	$7500 - 76.93 \cdot 75 = 1730.25$	

3.2 Internal Collaboration

Advances in information technology and enterprise systems have increased the availability of real-time data. This, in turn, has led to increased levels of information sharing and collaboration among the divisions (or business units) of a company. Before 1997, each division of Siemens Medical Systems had its own supplier and this significantly deteriorated buying power. Centralization of purchasing has saved 25% on material costs (Carbone (2001)). Similarly, Chevron Corp. is aiming to cut 5% to 15% from annual expenditures by centralizing the procurement system and by leveraging volume buys (Reilly (2001)).

In this section we assume that the procurement function is centralized within each company. Equivalently, multiple divisions within a company collaborate for procurement. We also assume that buyers know the structure of the transportation cost *(ftc* and *vtc)* for each supplier, possibly specified by a long-term contract. Examples of such practices are commonly found in the procurement of transportation services, such as trucking or sea cargo. Shippers contract with multiple carriers where each contract specifies a volume-based price and capacity availability. However, the buyers usually do not have to commit to a shipment volume in these contracts; even if they do, such minimum volume commitments are typically not enforced by the carriers. In our model, the price structure is defined by fixed and variable costs, i.e., suppliers offer volume discounts. We assume that a buyer i can request and receive information about the available total capacity, cap_k, and the volume-based price quote $P_k(d_{ij})$ (first two terms of equation (3.1)) for any item j from a supplier k for her entire demand d_{ij}. This implies that a buyer can determine the total cost for an order using the information on the transportation cost component and the quote on the production cost component for any item-supplier combination.

Under internal collaboration each buyer decides how much of each product to procure from each supplier using a centralized mechanism. The procurement decision of buyer i can be modelled by the following linear integer program.

x_{ijk}: 1, if buyer i contracts with supplier k for item j.

y_{ik}: 1, if buyer i contracts with supplier k.

$$\max \sum_j \sum_k res_{ij} \cdot x_{ijk} - \sum_j \sum_k (P_k(d_{ij}) + vtc_{ik}) \cdot d_{ij} \cdot x_{ijk} - \sum_k ftc_{ik} \cdot y_{ik} \tag{3.5}$$

subject to

$$\sum_j c_{jk} \cdot d_{ij} \cdot x_{ijk} \leq cap_k \qquad \forall k \qquad (3.6)$$

(IP-I)

$$x_{ijk} \leq y_{ik} \qquad \forall j, k \qquad (3.7)$$

$$\sum_k x_{ijk} = 1 \qquad \forall j \qquad (3.8)$$

Constraints (3.6) ensure that the amount of demand satisfied by a supplier does not exceed the current available capacity of the supplier. Constraints (3.7) ensure that when a contract is made with a supplier the corresponding fixed transportation cost is charged to the buyer. Constraints (3.8) ensure that the buyer contracts with a single supplier per item. (Note that a more compact formulation is possible by multiplying the right hand side of (3.6) by y_{ik} and removing constraints (3.7).)

The buyers contract with the suppliers sequentially as in the no collaboration model. Constraints (3.8) imply that even if the buyer surplus is negative after solving IP-I, the buyer still makes the contract. Each buyer makes the contracting decision based on maximizing her current surplus in equation (3.5). In the final matching, the contract price and the surplus of buyer i for item j is obtained as in equations (3.3) and (3.4).

Example (cont.): We analyze our previous example assuming that each company makes its procurement decisions centrally. Buyer 1 (both divisions of company 1) arrives to the market and makes the contracting decisions for both items by solving IP-I. Based on the outcome of the model, buyer 1 contracts with supplier 2 for both items and is charged the fixed transportation cost only once.

When buyer 1 makes her contract, supplier 2 initiates production for both items. Therefore, if buyer 2 also contracts with supplier 2, the associated fixed production cost for both items will be shared among the two buyers. Buyer 2 solves the same model with updated $P_2(d_{21})$ and $P_2(d_{22})$ values. The solution suggests that buyer 2 also contracts with supplier 2 for both items. Therefore the buyers benefit from both economies of scale and scope.

For buyer 1 the final contract price per unit of item 1 is 87.26 and the final price per unit of item 2 is 85.46. Therefore the total surplus of buyer 1 is 123.10. For buyer 2, the contract prices per unit of item 1 and 2 are, 71.57 and 69.76. The total surplus of buyer 2 is 2052.30. As compared to the traditional market, both buyers have increased their total surplus. The overall buyer surplus obtained in the market is 2175.40.

Table 3.7. Contract prices and surplus

	I_1	I_2	$surplus_i$
B_1	$price_{112} = 87.26$	$price_{122} = 85.46$	$(2000 - 87.26 \cdot 40)$ $+(3750 - 85.46 \cdot 25) = 123.10$
B_2	$price_{212} = 71.57$	$price_{222} = 69.76$	$(500 - 71.57 \cdot 10)$ $+(7500 - 69.76 \cdot 75) = 2052.30$

3.3 Full Collaboration

In the full collaboration model, we assume that a third party intermediary enables collaboration among multiple buyers. We use the terms *e-market* and *full collaboration* interchangeably. However, in practice not all e-markets enable full collaboration. Some e-markets provide only catalog services where suppliers and buyers post supply and demand quantities. 3PL providers such as Transplace or C.H. Robinson, where shippers and carriers do not contract directly with each other but through the 3PL intermediary, may enable full collaboration.

In this model, each buyer submits her demand and reservation price for each item, and each supplier submits cost and capacity information to the intermediary. The intermediary in turn matches supply and demand in the market with the objective of maximizing the total buyer surplus.

The matching program faced by the intermediary can be modeled by the following linear integer problem, which is a slight modification of (IP-I):

z_{jk}: 1, if supplier k initiates production for item j.

$$\max \sum_i \sum_j \sum_k res_{ij} \cdot x_{ijk} - \sum_j \sum_k fpc_{jk} \cdot z_{jk} - \sum_i \sum_k ftc_{ik} \cdot y_{ik}$$
$$- \sum_i \sum_j \sum_k (vpc_{jk} + vtc_{ik}) \cdot d_{ij} \cdot x_{ijk} \quad (3.9)$$

subject to

$$(3.7) \text{ and } (3.8) \qquad \forall i$$

$$\sum_i \sum_j c_{jk} \cdot d_{ij} \cdot x_{ijk} \leq tc_k \qquad \forall k \qquad (3.10)$$

$$x_{ijk} \leq z_{jk} \qquad \forall i, j, k \qquad (3.11)$$

Constraint (3.10) ensures that the amount of demand satisfied by a supplier does not exceed the total capacity of the supplier. Constraint (3.11) ensures that when production is initiated at a supplier for an item, a fixed production cost is incurred for that item.

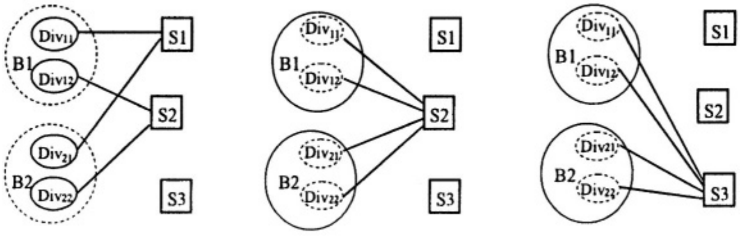
In the final matching, the contract price and surplus of buyer i for item j is obtained as in equations (3.3) and (3.4). While matching supply and demand, it is possible that some buyers have a negative surplus.

Example (cont.): Using our previous example, we illustrate the full collaboration model where the intermediary simultaneously matches buyers and suppliers. Under this model both companies contract with supplier 3 on both items. The contract prices and the surplus are given in Table 3.8. The total buyer surplus under this model is 6685.00. See Table 3.9 for a comparison of the surplus obtained by each buyer under three collaboration models.

Table 3.8. Contract prices and surplus

	I_1	I_2	$surplus_i$
B_1	$price_{113} = 17.92$	$price_{123} = 61.97$	$(2000 - 17.92 \cdot 40)$ $+(3750 - 61.97 \cdot 25) = 3483.75$
B_2	$price_{213} = 17.58$	$price_{223} = 61.63$	$(500 - 17.58 \cdot 10)$ $+(7500 - 61.63 \cdot 75) = 3201.25$

As seen in Figure 3.2, increasing collaboration levels among buyers leads to different matchings.



(a) No collaboration (b) Internal collaboration (c) Full collaboration

Figure 3.2. Matchings under different models of buyer collaboration

In these collaborative environments, the intermediary should well assess the implications of antitrust laws. Collaboration in the e-market might give incentives to the participants to collude in the upstream or downstream marketplaces. Therefore intermediaries should keep sensitive information such as output levels, reservation prices, costs or capacity levels, confidential. The total surplus obtained under full collaboration provides an upper bound on the surplus that would be obtained under less collaborative environments. Thus even if full collaboration is not possible due to antitrust laws or other reasons, the upper bound would provide valuable information for the intermediary and the participants.

Table 3.9. Surplus obtained by each buyer

	B_1	B_2	Total
No collaboration	105.00	2010.25	2115.25
Internal collaboration	123.10	2052.30	2175.40
Full collaboration	3483.75	3201.25	6685.00

4. Experimental Design

In this section we test how the three collaboration models perform under different market conditions. We consider a market with 25 buyers, 6 suppliers and 3 items. Each buyer places RFQs for all 3 items and suppliers have the capability to produce any of the items. We assume that the capacity required to produce one unit of any item is 1.

The parameters that define the marketplace are listed in Table 3.10. We selected the following three factors for controlling the market structure:

1 Market supply (total capacity).

2 Manufacturing set-up cost (fpc) versus variable production cost (vpc).

3 Fixed transportation cost (ftc) versus variable transportation cost (vtc).

We define two levels (low and high) for each of the three factors and obtain 8 different market settings.

4.1 Design Parameters

4.1.1 Market Supply vs. Market Demand. The demand of each buyer for each item is generated from a uniform distribution U[$\underline{d}$,$\overline{d}$]

(see Table 3.10). The market supply is defined as the total production capacity of the market, which can be either "low" or "high" compared to the expected total demand in the market. We model low (high) market supply by setting the supply equal to the lower (upper) bound of the total demand.

Average total capacity required to satisfy low market demand =

$$TC_{low} = (\#\text{ of buyers}) \cdot (\#\text{ of items per buyer}) \cdot (\underline{d}\text{ per item})$$
$$\cdot(\text{avg. capacity required per unit})$$

Average total capacity required to satisfy high market demand =

$$TC_{high} = (\#\text{ of buyers}) \cdot (\#\text{ of items per buyer}) \cdot (\bar{d}\text{ per item})$$
$$\cdot(\text{avg. capacity required per unit})$$

In our experiments we assume that the capacities of the suppliers are equal and hence the capacity per supplier is the total market capacity divided by the number of suppliers.

Table 3.10. The parameters of the market

Demand per item	$U(10,20)$
Total capacity per supplier	125 for tight, 250 for relaxed
fpc	$U(46,54)$ $U(980,1020)$
vpc	$U(8,12)$
Reservation price per item	$U(300,370)$ $U(300,1320)$ $U(300,850)$ $U(300,1800)$
ftc	$U(18,22)$ $U(485,515)$
vtc	$U(4,6)$

4.1.2 Fixed Costs of Manufacturing and Transportation.

The average setup cost of production, i.e., the mean of fpc, is set at either "low" or "high" levels as compared to the average variable production cost vpc, where μ_{fpc}=50 or 1000 and μ_{vpc}=10. fpc values are generated randomly from the uniform distributions U[46,54] and U[980,1020], for the low and high fpc levels, respectively.

The fixed cost of transportation, ftc, is set either at "low" or "high" levels as compared to vtc, where μ_{ftc}=20 or 500 and μ_{vtc}=5. ftc values are generated randomly from the uniform distributions U[18,22] and U[485,515], for the low and high cases, respectively.

Note that rather than the individual values of fpc, ftc, vpc and vtc, we consider the ratios $\frac{fpc}{vpc}$ and $\frac{ftc}{vtc}$ as the design factors in our experiments.

4.1.3 Reservation Price. We generate the reservation prices of the buyers randomly from the uniform distribution $U \sim [(vpc + vtc) \cdot \underline{d}, fpc + ftc + (vpc + vtc) \cdot \overline{d}]$. The lower bound corresponds to the case where (in the limit) the buyer only pays the variable costs. The upper bound corresponds to the worst case where a buyer incurs the entire fixed production and transportation cost, in addition to the variable costs.

Table 3.11. Upper and lower bound for reservation price in four different market types

Market type	μ_{fpc}, μ_{ftc}	Reservation price
1	(50,20)	U(300,370)
2	(1000,20)	U(300,1320)
3	(50,500)	U(300,850)
4	(1000,500)	U(300,1800)

The design factors and factor levels are shown in Table 3.12.

Table 3.12. Design factors and factor levels

		Levels	
		low	high
Factor 1	Supply/Demand	.66	1.33
Factor 2	fpc/vpc	5	100
Factor 3	ftc/vtc	4	10

4.2 Buyer strategies

In our experiments, we consider the following buyer strategies defined previously:

1 No collaboration

  (a) Myopic strategy (MYOPIC)

  (b) Accept the minimum bid (MIN)

  (c) Leave and possibly return later (LOQ) with $\beta = \frac{D}{Q}$

  (d) Leave and return later (Q)

  (e) Accept the lowest bid or leave and return later (AOQ) with $\alpha = 0.25$

  (f) Minimum optimistic bid (MOB)

2 Internal collaboration (INT)

3 Full collaboration (e-MARKET)

4.3 Performance Measures

To evaluate the effectiveness of different bidding strategies under various market conditions, we consider the following performance measures:

1 % of satisfied demand $= \frac{\text{quantity of satisfied demand}}{\text{total market demand}} \cdot 100$

2 total surplus = total reservation price - total cost

3 average surplus per unit$= \frac{\text{total surplus}}{\text{quantity of satisfied demand}}$

4 average cost per unit$= \frac{\text{total cost}}{\text{quantity of satisfied demand}}$

From a buyer's perspective, it is desirable to have a high amount of satisfied demand (market liquidity), and a high surplus per unit of satisfied demand (quality of surplus), leading to a high total buyer surplus. Note that the quality of surplus is inversely proportional to the suppliers' costs per unit, i.e., a strategy which decreases the setup effort in the market is desirable.

Current e-marketplaces focus on similar performance measures. For example, Transplace uses the capacity of several carriers to satisfy demand, which makes it an attractive choice for shippers in terms of demand satisfaction. Similarly, consortia e-marketplaces such as `covisint.com` help buyers to achieve economies of scale and therefore reduce average cost per unit.

5. Experimental Results

In this section we compare the strategies with respect to the performance measures. Each market consists of 25 buyers, 6 suppliers, and 3 items. The results are obtained and compared using a t-test for 8 market types and 8 strategies, with 15 runs for each market type and strategy combination.

The market types are indicated by a 3 digit code abc, where $a,b,c \in \{0,1\}$. 0 corresponds to *low* level setting for that factor and 1 corresponds to *high* level. a indicates capacity level in the market, b indicates $\frac{fpc}{vpc}$ level and c indicates $\frac{ftc}{vtc}$ level. For example, the 101 market corresponds to high capacity, low fixed cost of production and high fixed cost of transportation.

The results are presented in Tables 3.13-3.19. In Tables 3.14, 3.16, 3.18 and 3.20 a ">" sign indicates that a strategy has resulted in a higher value at the indicated significance level (SL). If two strategies are in the same set, then their performances are not significantly different from each other.

5.1 Results for the percentage of satisfied demand

OBSERVATION 3.1 *In maximizing the percentage of satisfied demand, when capacity is high*

i. AOQ, MIN, MOB *and* INT *perform best, followed by e-*MARKET, *followed by* Q, *followed by* LOQ *and* MYOPIC.

ii. If there are no economies of scale and scope (market 100), then LOQ *and* MYOPIC *have similar performance; otherwise,* LOQ *outperforms* MYOPIC.

It is not surprising that MOB, MIN and INT perform well in satisfying the demand since they do not consider reservation prices. Since AOQ accepts offers with some probability even if they are above the reservation price, it also performs well in satisfying the demand.

OBSERVATION 3.2 *In maximizing the percentage of satisfied demand, when capacity is low*

*i. e-*MARKET *is at least as good as any other strategy.*

ii. When benefits from economies of scale or scope are high MYOPIC *and* LOQ *are always the worst, followed by* Q.

iii. MIN, INT, AOQ *and* MOB *have similar performance when there* *are economies of scope. When there are no economies of scope,* e-MARKET *dominates other strategies.*

Note that MYOPIC and LOQ are always the worst in satisfying the demand except for 000, followed by Q, regardless of the capacity level in the market.

Table 3.13. Satisfied demand in different markets (all figures are percentages)

Market type	MYOPIC	MIN	Q	AOQ	LOQ	MOB	INT	e-MARKET
000	65.0	65.0	65.0	64.8	65.0	64.3	64.8	65.5
001	21.2	64.5	27.2	64.8	21.2	64.5	64.8	64.9
010	29.8	64.9	35.9	65.0	29.8	63.9	64.6	65.4
011	19.9	64.5	26.6	64.5	19.9	64.5	64.6	65.1
100	95.1	100.0	97.4	100.0	95.5	100.0	100.0	99.4
101	22.2	100.0	35.4	99.8	25.0	100.0	100.0	94.7
110	40.9	100.0	57.1	100.0	46.7	100.0	100.0	96.3
111	26.0	100.0	42.7	99.8	30.2	100.0	100.0	84.7

Table 3.14. Comparison of strategies with respect to the percentage of satisfied demand

Market type	Performance	SL
000	e-MARKET > {MIN, MYOPIC=LOQ=Q, AOQ,INT, MOB}	97%
001	{e-MARKET, INT, AOQ, MIN, MOB} > Q > {MYOPIC= LOQ}	99%
010	e-MARKET > {AOQ, MIN } > INT > MOB > Q > {MYOPIC=LOQ}	96%
011	{ e-MARKET, INT, MIN,MOB, AOQ} > Q >{MYOPIC=LOQ}	99%
100	{AOQ=MIN=MOB=INT} > e-MARKET > Q > {LOQ,MYOPIC}	99%
101	{MIN=MOB=INT, AOQ, }> e-MARKET > Q > LOQ > MYOPIC	99%
110	{AOQ=MIN=MOB=INT}> e-MARKET > Q > LOQ > MYOPIC	99%
111	{MIN=MOB= INT, AOQ,}> e-MARKET > Q > LOQ > MYOPIC	99%

5.2 Results for total surplus

OBSERVATION 3.3 *e-*MARKET *outperforms all other strategies in terms of total surplus under any market structure. The benefit of e-*MARKET *compared to the next best strategy is highest when the capacity is low and there are economies of scale and scope (Figure 3.3).*

OBSERVATION 3.4 *In maximizing the total surplus*

i. INT *outperforms* MOB *when there are economies of scope (markets 001, 011, 101 and 111).*

ii. MIN *outperforms* INT *when there are economies of scale but not scope (markets 010 and 110).*

iii. Q *outperforms* LOQ *when there are either economies of scale or scope or when the capacity is high. In market 000,* Q *and* LOQ *have the same performance.*

iv. MOB *outperforms or does as well as* MIN *in all markets. The difference between* MOB *and* MIN *is greatest when the capacity is high and there are only economies of scope (market 101).*

v. AOQ *outperforms* LOQ *when there are economies of scale (markets 010, 011, 110, 111); otherwise* LOQ *either outperforms or does as well as* AOQ.

Observation 3.4.*i* is in line with the fact that INT tries to consolidate orders of the same buyer for different products and therefore saves on transportation cost, whereas MOB focuses on lowering the production cost by consolidating orders from different buyers for the same product.

Observation *3A.ii* implies that when the benefits from economies of scale are high, the MIN strategy leads to fewer production initiations as compared to the INT strategy. In the INT strategy, due to collaboration, the buyer divisions arrive to the system at the same time and contract with a set of suppliers to maximize their overall surplus. Although the fixed cost of transportation is low, experimental results indicate that it is still more beneficial for a buyer to contract with fewer suppliers than the number of items she demands. Please note that this may not be the case if the variance of *variable transportation cost* is sufficiently high. In that case a buyer could select a different supplier for each of her items. In the MIN strategy, buyer divisions arrive to the market independently. For most of the cases two divisions of the same buyer are not assigned to the same supplier because they arrive at different times and the same supplier is no longer available for the division that arrives later. Therefore in most of the instances a supplier uses all of his capacity to produce a single item, whereas in the INT strategy a supplier uses his capacity to produce two or more items. In the INT strategy, each buyer contracting with a supplier for two or more items causes more suppliers to initiate production as compared to the MIN strategy. As a result, while losses from economies of scale are high, gains from economies of scope are insignificant in the INT strategy.

Observation 3.4.*iv* indicates that the "lookahead" policy employed by the MOB strategy helps to increase the surplus compared to the MIN strategy, which does not consider potential future arrivals. Observations 3.4.*iii* and 3.4.*v* show that the total surplus can increase if the buyers accept the lowest bid even if it is above their reservation price, or return to the market with a positive probability.

We should note that the observations may partly depend on the experimental settings. For instance, observation 3.4.*ii* might change if the cost values assigned to each supplier had much smaller variance. In that case, the INT strategy would not necessarily lead to more production initiations and buyer-supplier assignments would have a similar structure to the MIN strategy.

In our experimental design we limited the number of items that each buyer is willing to buy to three, due to computational difficulties. However we conjecture that as the number of items increases, INT strategy will achieve a higher total surplus in the presence of economies of scope. When only economies of scale exist, INT strategy might perform worse due to observation 3.4.*ii*.

Table 3.15. Total surplus in different markets (all figures are in thousands).

Market type	MYOPIC	MIN	Q	AOQ	LOQ	MOB	INT	e-MKT
000	5.031	4.629	5.032	4.940	4.503	4.530	4.588	6.703
001	3.220	−3.495	4.614	−3.142	3.220	−1.850	6.046	10.706
010	13.404	16.744	16.176	21.444	13.404	18.245	13.681	29.635
011	8.418	5.064	11.217	12.826	8.418	6.653	10.108	24.932
100	7.607	7.617	7.647	7.626	7.611	7.612	7.750	8.318
101	3.557	7.002	6.671	3.693	4.343	5.970	12.555	13.078
110	18.201	33.263	24.332	33.655	20.330	34.265	31.781	35.950
111	11.724	16.597	18.110	16.768	13.367	20.300	30.563	33.502

5.3 Results for average surplus per unit

OBSERVATION 3.5 *In maximizing the average surplus, when the capacity is low*

i. *e*-MARKET *outperforms or does at least as well as any other strategy.*

ii. AOQ *outperforms* MOB *and* MIN.

Table 3.16. Comparison of the strategies with respect to total surplus in different markets

Market type	Performance	SL
000	e-MARKET > {MYOPIC=LOQ=Q} > AOQ > {MIN, INT, MOB}	99%
001	e-MARKET > INT > Q > {MYOPIC=LOQ} > AOQ > MOB > MIN	99%
010	e-MARKET > AOQ > MOB > {MIN, Q} > {INT, MYOPIC=LOQ}	95%
011	e-MARKET > AOQ > {Q,INT} > {MYOPIC=LOQ} >MOB > MIN	90%
100	e-MARKET > INT > Q >{AOQ, MIN, MOB, LOQ, MYOPIC}	90%
101	e-MARKET > INT > {Q,MOB} > {LOQ, AOQ, MYOPIC}> MIN	98%
110	e-MARKET > {MOB, AOQ,MIN} > INT > Q > LOQ> MYOPIC	99%
111	e-MARKET> INT > {MOB,AOQ, Q, MIN} > {LOQ,MYOPIC}	99%

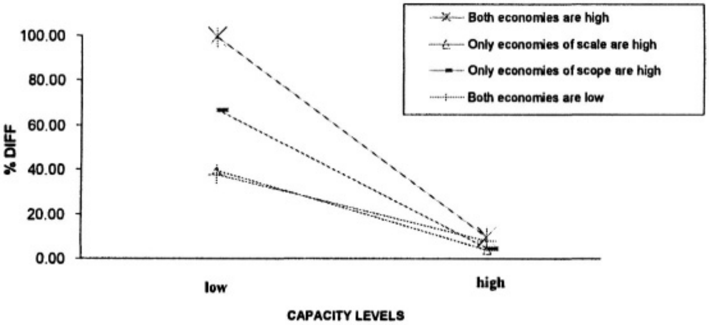


Figure 3.3. % difference in the total surplus between e-MARKET and the next best strategy under different market types

OBSERVATION 3.6 *In maximizing the average surplus,*

i. MYOPIC, Q *and* LOQ *perform at least as well as or better than all strategies except e-*MARKET.

ii. MIN *has the worst performance except when there are only economies of scale (markets 010 and 110), in which case* INT *performs worst.*

Note that the strategies which consider reservation prices (MYOPIC, LOQ or Q) result in higher average surplus compared to the strategies which do not. Recall that these strategies do not perform well in satisfying the demand (Observation 3.1). On the other hand, strategies which do not consider reservation prices (MIN and MOB) result in lower average surplus levels, but higher percentages of satisfied demand. In these strategies it is possible that some buyers obtain a negative surplus. These results imply a significant trade-off between high satisfied demand and the average surplus.

Under tight capacity Observation 3.4.*i* also holds for maximizing the average surplus, i.e., MOB does better when there is economies of scale and INT does better when there are economies of scope.

Table 3.17. **Average surplus per unit in different markets**

Market type	MYOPIC	MIN	Q	AOQ	LOQ	MOB	INT	e-MARKET
000	6.9	6.3	6.9	6.8	6.9	6.3	6.3	9.1
001	13.5	−4.8	15.3	−0.4	13.5	−2.5	8.3	14.7
010	40.1	22.9	40.3	29.3	40.1	25.4	18.9	40.3
011	37.3	7.0	37.5	17.7	37.3	9.2	13.9	34.1
100	7.1	6.8	7.0	6.8	7.1	6.8	6.9	7.5
101	14.1	0.6	16.8	3.0	15.4	5.3	11.2	12.3
110	39.5	29.6	38.3	29.9	38.7	30.5	28.3	33.3
111	40.2	14.8	38.5	15.0	39.5	18.1	27.2	35.4

5.4 Results for average cost per unit

OBSERVATION 3.7 *e-*MARKET *always has the lowest average cost, except when the capacity is high and there are only economies of scale (market 110).*

OBSERVATION 3.8 MIN *has the highest average cost except when there are only economies of scale (markets 010 and 110), in which case* INT *has the highest average cost.*

Table 3.18. Comparison of the strategies with respect to average surplus in different markets

Market type	Performance	SL
000	e-MARKET > {MYOPIC=LOQ=Q} > AOQ >{MIN, INT, MOB}	99%
001	{Q, e-MARKET} >{MYOPIC=LOQ} > INT >AOQ> MOB > MIN	88%
010	{e-MARKET, Q,MYOPIC=LOQ} > AOQ > MOB> MIN > INT	99%
011	{Q, MYOPIC=LOQ,e-MARKET} > AOQ > INT >MOB > MIN	99%
100	e-MARKET > MYOPIC > LOQ >{Q, INT} > {AOQ, MIN,MOB}	90%
101	Q > LOQ > MYOPIC >e-MARKET > INT > MOB >AOQ > MIN	95%
110	MYOPIC > {LOQ, Q} >e-MARKET > {MOB, AOQ, MIN}> INT	95%
111	{MYOPIC, LOQ, Q} >e-MARKET > INT > MOB >{AOQ, MIN}	93%

It is interesting to note that when INT is not the worst performer in average cost, it is the second best. This leads us to conclude that collaboration helps to decrease the average cost per unit, especially when economies of scale and scope are high. In comparing MOB and INT, Observation 3.4.*i* also holds for minimizing the average cost, i.e., MOB does better when there are economies of scale and INT does better when there are economies of scope.

Table 3.19. Average cost per unit in different markets

Market type	MYOPIC	MIN	Q	AOQ	LOQ	MOB	INT	e-MARKET
000	16.0	16.2	16.0	16.0	16.0	16.2	16.0	15.4
001	40.8	42.6	36.5	41.3	40.8	40.5	29.4	28.3
010	26.8	29.8	24.7	28.7	26.8	27.5	34.0	23.7
011	62.7	61.6	58.8	60.8	62.7	59.6	54.4	50.4
100	15.4	15.5	15.4	15.4	15.4	15.4	15.3	14.8
101	39.8	37.0	32.3	34.3	37.1	32.3	26.4	25.0
110	22.5	23.0	20.1	22.7	21.2	22.1	24.3	21.1
111	57.0	53.1	52.1	53.0	55.2	49.9	40.7	39.6

Before closing this section, we would like to briefly discuss how our assumptions of "the buyers being located in the same region" and "each buyer division being responsible for the procurement of one item" affect the experimental results. If these two assumptions were to be relaxed, then the structure of the model and the form of collaboration would change. In this case divisions that belong to different buyers

Table 3.20. Comparison of the strategies with respect to average cost in different markets

Market type	Performance	SL
000	{MOB, MIN} > {AOQ, Q=MYOPIC=LOQ, INT} > e-MARKET	99%
001	MIN > {AOQ,MYOPIC=LOQ, MOB} > Q >INT > e-MARKET	99%
010	INT > {MIN, AOQ} > {MOB, MYOPIC=LOQ} > Q > e-MARKET	95%
011	{MYOPIC=LOQ, MIN, AOQ}> {MOB, Q} > INT >e-MARKET	95%
100	{MOB, MIN, AOQ, Q,LOQ, MYOPIC, INT} >e-MARKET	99%
101	MYOPIC> {LOQ, MIN} >AOQ > {Q, MOB} > {INT,e-MARKET}	98%
110	INT > {MIN, AOQ,MYOPIC, MOB} > {LOQ,e-MARKET} > Q	80%
111	MYOPIC > LOQ > {MIN,AOQ, Q} > MOB > INT >e-MARKET	97%

located in the same region could achieve savings from transportation, whereas buyer divisions of the same buyer located at different regions could achieve savings from production by leveraging their purchasing power. This might lead to a conclusion that internal collaboration is beneficial when benefits from economies of scale are high. However, full collaboration would still be most beneficial when savings both from economies of scale and scope are high.

6. Conclusion and Future Work

We analyzed markets where multi-unit transactions over multiple items take place. We considered three different trading models with increasing levels of collaboration among buyers. The "no collaboration" model considers traditional markets where there is no collaboration among buyers or buyer divisions. In the "internal collaboration" model, purchasing divisions of a buyer collaborate for procurement. In the "full collaboration" model an intermediary enables collaboration among different buyers.

We studied six different buyer strategies for the no collaboration model, and one for the internal collaboration model. These strategies were tested against the centralized buyer-seller matching mechanism employed by the intermediary in the "full collaboration" model.

The experimental results show that when there is tight capacity in the market and when potential economies of scope are high (i.e., when the fixed cost of transportation is high), the "full collaboration" model results in significantly higher total surplus than the other strategies. The increase in surplus is even more pronounced when economies of scale are also high (i.e., when the fixed costs of both manufacturing and transportation are high). The extra benefits obtained by full collaboration are relatively low when the capacity is high and the fixed cost factors are low.

We also observe that internal collaboration performs very well, provided that the potential benefits from economies of scope are high. On the other hand, when the potential benefits from economies of scale are high, buyer strategies with a "look-ahead" perform well. These are the strategies which consider potential future trades in the market by other buyers while contracting with a supplier.

Our analysis indicates that the potential benefits of intermediaries are highest in capacitated markets with high fixed production and/or transportation costs. Process industries such as rubber, plastic, steel and paper are typically characterized by high fixed production costs. High fixed transportation costs can arise in industries that have to either manage their own distribution fleet or establish contracts with carriers that guarantee a minimum capacity in order to ensure adequate service levels (e.g., Ford Customer Service Division guarantees routed deliveries to their dealers every three days and hence has high fixed transportation costs. Due to variability in the demand that they face, it can be the case that there is fairly low capacity utilization).

It is important to note that different collaboration models require different coordination, personnel, and technology costs. Although our current models do not account for any fixed or variable costs of implementing a collaboration strategy, they could easily be modified to incorporate such information. Clearly, the benefits should outweigh the costs of implementing a particular strategy for that strategy to be attractive for a firm.

In our future work we are planning to extend the experimental analysis to a larger market size in an effort to observe the effect of market liquidity on performance. In addition, we would like to design allocation mechanisms that are both computationally efficient and perform well across a broad range of performance measures. Finally, it would be interesting to consider the issue of pricing strategies for the intermediary such as charging a subscription fee to buyers and/or suppliers versus charging the participants for each transaction.

Acknowledgements

Pinar Keskinocak is supported by NSF Career Award DMI-0093844. This research is also supported in part by NSF grant EEL-9802942. The authors are grateful to an anonymous referee and Joseph Geunes for their constructive feedback throughout the revision of this chapter.

References

Aviv, Y. 2001. The Effect of Collaborative Forecasting on Supply Chain Performance. *Management Science* 47, 1326–1343.

Cachon, G.P. 2003. *Supply Chain Coordination with Contracts, S. Graves, T. de Kok, eds., Supply Chain Management-Handbook in OR/MS (forthcoming).* North-Holland, Amsterdam.

Cachon, G.P. and Zipkin, P.H. 1999. Competitive and Cooperative Inventory policies in a Two-Stage Supply Chain. *Management Science* 45, 936–953.

Carbone, J. 2001. Strategic Purchasing Cuts Costs 25% at Siemens. *Purchasing,* September 20, 29–34.

Cetinkaya, S. and Lee, C.Y. 2000. Stock Replenishment and Shipment Scheduling for Vendor-Managed Inventory Systems. *Management Science* 46, 217–232.

Cheung, K.L. and Lee, H.L. 2002. The Inventory Benefit of Shipment Coordination and Stock Rebalancing in a Supply Chain. *Management Science* 48, 300–306.

Elmaghraby, W. and Keskinocak, P. 2002. Ownership in Digital Marketplaces. *European American Business Journal,* 71–74.

Gurnani, H. 2001. A Study of Quantity discount Pricing Models with Different Ordering Structures: Order Coordination, Order Consolidation, and Multi-Tier Ordering Hierarchy. *International Journal of Production Economics* 72, 203–225.

Jin, M. and Wu, D. 2001. Supply Chain Contracting in Electronic Markets: Auction and Contracting Mechanisms. Working Paper, Lehigh University, Bethlehem, PA.

Kalagnanam, J., Davenport, A.J. and Lee, H.S. 2001. Computational Aspects of Clearing Continuous Call Double Auctions with Assignment Constraints and Indivisible Demand. *Electronic Commerce Research* 1, 221–238.

Keenan, F. and Ante, S.E. 2002. The New Teamwork. *Business Week e.biz,* February 18, 12–16.

Keskinocak, P. and Tayur, S. 2001. Quantitative Analysis for Internet-Enabled Supply Chains. *Interfaces* 31, 70–89.

Ledyard, J.O., Banks, J.S. and Porter, D.P. 1989. Allocation of Unresponsive and Uncertain Resources. *RAND Journal of Economics* 20, 1–25.

Lee, H.L., Padmanbhan, V., Taylor, T.A. and Whang, S. 2000. Price Protection in the Personal Computer Industry. *Management Science* 46, 467–482.

Reilly, C. 2001. Chevron Restructures to Leverage Its Buying Volumes. *Purchasing,* August 9.

Reilly, C. 2002. Specialists Leverage Surfactants, Plastics, Indirect and Other Products. *Purchasing,* January 15.

Smyrlis, L. 2000. Secrets of a Successful Marriage: 3PL Users Want a Deeper, more Meaningful Relationship with their 3PL Service Providers. Are 3PLs ready to commit? *Canadian Transportation Logistics,* February.

Stackpole, B. 2001. Software Leaders - Trading Exchange Platforms. *Managing Automation,* June.

Strozniak, P. 2001. Sharing the Load. *IndustryWeek,* September 1, 47–51.

Weng, Z.K. 1995. Channel Coordination and Quantity Discounts. *Management Science* 41, 1509–1522.

Chapter 4

THE IMPACT OF NEW SUPPLY CHAIN MANAGEMENT PRACTICES ON THE DECISION TOOLS REQUIRED BY THE TRUCKING INDUSTRY

Jacques Roy
Production and Operations Management Department
HEC Montréal
Montréal, Canada
jacques.roy@hec.ca

Abstract For the last twenty years or so, the freight transportation industry has been facing new challenges such as time-sensitive industrial and commercial practices as well as the globalization of markets. During the same period, new information-related technologies have developed rapidly: Electronic Data Interchange (EDI) and the Internet, Global Positioning Systems via satellites (GPS) and Decision Support Systems (DSS). These technologies can greatly enhance the operations planning capability of freight carriers in as much as they make use of this information in order to optimize their operations. GPS, EDI and the Internet can also provide the necessary information required to achieve real-time computer-based decision making using appropriate operations research techniques. Today's decision support tools must therefore be designed to be used in a real-time environment. We describe this environment and propose optimization tools that can be made available to motor carriers.

1. Introduction

During the last twenty years or so, the freight transportation industry had to face new challenges as a result of important changes affecting supply chains and logistical processes. The first change may be attributed to the impetus toward inventory reduction which led to *Just-In-time* procurement practices and, more recently, to *Quick Response* or *Efficient Consumer Response,* that is the just-in-time replenishment of goods in the retail and grocery industries. The procurement and distribution of goods has also been significantly influenced by the recent trend toward the globalization and liberalization of markets. This has led to free trade agreements between countries over wide geographical regions such as the European Union and the North American Free Trade Agreement (NAFTA). These changes have resulted in increased demands for electronic trade and specialized logistical services such as *third party providers.* In addition, with the advent of electronic commerce, *Business-to-Business* and *Business-to-Consumer* practices are now very much part of the motor carrier environment.

As a result of these changes, the freight transportation industry has had to recognize the importance of information technologies in order to enhance its capability to respond to the needs of its customers. New information related technologies such as electronic data interchange (EDI), Internet, and global positioning systems (GPS) via satellites have been adopted by most of the leading motor carriers during the last decade or so. These technologies can greatly enhance the planning capability of freight carriers in as much as they make use of this information in order to optimize their operations. GPS, EDI, and the Internet can provide the necessary information required to perform real-time dispatch using appropriate operations research techniques. With on-board computers, motor carriers can be informed of the content of pick-up vehicles before they reach the terminal. They can therefore decide the best door assignment rule for these trucks in order to minimize freight handling. They can also revise work scheduling and, more importantly, the load planning of highway vehicles on a daily basis, thus generating major savings while providing high levels of service to customers. Several of these tools were initially developed as tactical or operational planning tools. Today's decision support tools must be designed to be used in a real-time environment.

This chapter begins with a review of recent trends in logistics and of their impact on motor carriers. The next section describes the typical operations of a motor carrier. Then, the more "traditional" tactical and operational planning tools are quickly reviewed, followed by a description of the set of real-time decision tools that are required in today's trucking industry.

2. Recent trends in supply chain management and their impact on the trucking industry

This section introduces some of the most important recent trends in logistics, or what is now more commonly named *Supply Chain Management,* and discusses their impact on the motor carrier industry.

2.1 Increased use of time-sensitive strategies in both manufacturing and commerce

The trend toward time-sensitive strategies originated in the 1970's with the advent of JIT practices in the manufacturing sector as a reaction to the successes recorded in Japan by companies such as Toyota. Its philosophy rests on the elimination of waste and focuses on the reduction of production set-up times and inventories in the supply chain. Although JIT practices have been around for several years now, their application is still growing today and it is estimated that the majority of shipments are currently ordered just-in-time in the United States (Bowersox and Taylor, 2001).

During the 1990's, the JIT philosophy extended to the physical distribution of finished goods from manufacturing facilities to retail outlets and through distribution centers. This has led to new practices known as *continuous replenishment programs* (CRP), *quick response* (QR) and *efficient consumer response* (ECR). CRP defines the practice of partnership between members of a distribution network that modifies the traditional approach of replenishment based on customer orders, to a replenishment strategy based on both the actual demand at the point-of-sale (POS) and sales forecast. The sharing of POS information with suppliers is a critical aspect of these new practices. Other key elements include: shorter reaction times; more efficient logistics networks based on faster transportation means, cross-docking and more efficient in-store reception methods; using EDI or the Internet in combination with bar coding; emphasis on total quality management and continuous improvement practices; activity based costing and category management. QR is the name given to such practices in the textile and apparel industry. Major savings have been reported by well-known companies such as Benetton, Sears Roebuck, Levi's, and so on. ECR is the grocery industry's answer to QR and it is expected to reduce costs by about 10.8 percent in the Canadian food industry. These new practices are expected to spread to other sectors as well; we are, for example, referring to *efficient healthcare consumer response* (EHCR).

Vendor-managed inventory (VMI) systems differ from CRP in that the vendor or manufacturer actually makes the decision on inventory policies and replenishments at the retailer level. In VMI, the supplier assumes responsibility for replenishing retail inventories in the required quantities and formats based on sales data provided by the retailer (Bowersox et al., 2002). VMI arrangements have been exemplified by Wal-Mart and Proctor & Gamble, Costco and Kimberly-Clark and others with reported increases in sales and inventory turnover in the order of 20 to 30 percent (Simchi-Levi et al., 2003). When inventories are managed cooperatively by sharing information between retailers and manufacturers, the practice is sometimes referred to as *co-managed inventory* (CMI). One of the first published experimentations of CMI in Europe was in the grocery retail sector, which is reported in Winter (1999). Finally, *collaborative planning, forecasting and replenishment* (CPFR) is a process that enables collaboration between retailers and suppliers. It is a tool that facilitates the exchange of data between supply chain partners for the establishment of forecasts, following an iterative process. With CPFR, a common forecast is created and shared between partners via EDI or the Internet in order to improve the planning process.

The impact of these time-sensitive strategies on motor carriers is twofold. First of all, they are making use of their terminal facilities to perform cross-docking of shipments between manufacturers and retailers, thereby increasingly replacing distributors in that role. Secondly, they are acting as partners in the distribution of goods between manufacturers and retailers. This partnership can mean the fulfillment of additional services such as the temporary storage of goods, inventory management and control, and other value added services such as the creation of assortments, packing and conditioning of products. To accomplish these activities, motor carriers must be linked to the real time information networks of their partners as we will see in the next section.

2.2 Development of electronic commerce

Electronic commerce is defined as conducting or enabling the buying and selling of goods or services through electronic networks, including the Internet (Gopal and Fox, 1996). It makes use of a wide array of information technologies such as fax, E-mail, voice mail, electronic funds transfer, the Internet, Intranet, image processing, barcode and EDI. The latter was certainly a pivotal point in the development of electronic commerce as pointed out by many authors (Rogers et al., 1993; Raney and Walter, 1992; Masters et al., 1991). These works all have shown the importance of EDI in making logistics and physical distribution more efficient and faster. More recently though, the Internet has appeared to be a better and cheaper way to get the same results (Dismukes and Godin,

1997) because cost and a lack of critical mass of traditional EDI users in the market were major barriers to EDI implementation, particularly for small carriers (Udo and Pickett, 1994, 1994). Moreover, investing in information technologies such as EDI may only be profitable when these technologies are fully integrated with other internal systems within the organization (Ratliff, 1995).

The use of those information and communication technologies thus appears to be a determinant in achieving excellence in logistics. Nearly a decade ago, the results of a survey of over 200 top managers of logistics and transportation in the United States were already pointing out information and communication technologies as the most important success factor in their domain (Masters et al., 1991). The World Class Logistics research completed at Michigan State University with its 3,693 respondents surveyed, identified information technology as one of the four crucial factors defining world class logistics capability. From a subset of respondents to this last research, the authors demonstrated more precisely the prevalence of this factor in determining logistics capability (Closs et al., 1997). The last annual Deloitte & Touche consulting group's survey of North American trends in supply chain management also finds, from a sample of nearly 200 manufacturers and distributors, that information technology is the key to supply chain success (Factor and Kilpatrick, 1998). Among the benefits associated with use of these technologies we can underline the following: minimization of manual data entry, increased transaction speed and accuracy, lower communication costs and promotion of simplification of procedures.

The use of electronic commerce also implies that all participants in a supply chain, including motor carriers, become essential to the process. It is the entire supply chain, from materials to finished goods, that is thus affected by the introduction of these technologies. In fact, electronic commerce acts both ways: as a response to a better supply chain performance and also as a source of pressure from the market to improve the performance of supply chains. QR and ECR are facets of what is now commonly referred to as business-to-business (B2B), which is the application of e-commerce between businesses. When consumers are interacting directly with suppliers, we are in a business-to-consumer (B2C) environment, where not only the distributors are being eliminated but the retailers as well! The role of freight carriers is increasingly important because the physical distribution of goods is the most important physical activity left in such an environment. Motor carriers are therefore under strong incentive to join the movement and integrate new information and communication technologies into their day-to-day operations.

The number of motor carriers offering EDI services in the U.S. and Canada has thus increased significantly in recent years (Sunstrum and Howard, 1996). However, Canadian motor carriers have been slower to adopt EDI and electronic commerce, as indicated in a survey of motor carriers using best practices conducted by KPMG in 1997. Golob and Regan (2002) report that 31.3 percent of the 1136 trucking companies surveyed in the state of California were using EDI in 1998. They also predict that EDI will most likely be adopted by the larger fleet operators and least likely by private fleets. Moreover, many surveys have shown that carriers are more reactive than proactive in that matter. For instance, it was found that carriers adopt EDI mainly to satisfy customer requirements, and after to increase customer service and remain competitive (Walton, 1994).

We also see a growing number of carriers concluding partnership agreements or alliances with shippers and other carriers as well. These alliances are facilitated by the use of new information technologies that are part of the electronic commerce arsenal. This pattern may lead to virtual (or electronic) integration between carriers and shippers. However, carriers that elect not to invest in these new technologies will not be able to provide value added services to their customers and will therefore lose market share. Actually, the rapid growth of e-commerce (B2B and B2C) may well benefit third party logistics providers (Golob and Regan, 2001).

2.3 Outsourcing of logistical services to third party providers

With the globalization of markets, companies are increasingly focusing on their core competencies, be it the manufacturing or assembly of goods (e.g., the automotive industry) or the design and marketing of products (e.g., Nike). Those core competencies seldom include the supply and distribution of goods, and these activities are therefore increasingly outsourced to third party organizations that specialize in logistics. The U.S. market for third party logistics (3PL) services was estimated at $10 billion in 1990 (Sink and Langley, Jr., 1997). It has grown to $25 billion in 1996 and to $56 billion in 2000 (Chow and Gritta, 2002). According to a survey of Fortune 500 manufacturers conducted in 1999, 65 percent of the respondents indicated that their company used 3PL services and that roughly one third of their logistics operating budgets will be given to 3PL within three years (Lieb and Peluso, 1999). The same survey indicates that the most frequently used 3PL services are related to transportation and warehousing.

It is therefore not surprising to observe that a good number of 3PL services providers originated from the transportation industry. Some of the larger ones are Ryder, Schneider, UPS, and FedEx. It seems that it is increasingly difficult for motor carriers, in North America as well as in Europe, to obtain satisfactory financial results by concentrating only on the transportation sector. An increasing number of trucking companies are finding it necessary to diversify their services in order to include some of the logistical services currently being offered by 3PL services providers. The trucking industry may well consist of three major segments: (i) those carriers who have chosen to diversify and compete with 3PL's, (ii) some carriers fortunate enough to exploit a niche (specialized equipment or focus on a specific product), and (iii) the vast majority of remaining carriers who will have to reduce costs in order to survive in an increasingly competitive market. Even those who try to differentiate themselves from others by offering superior service levels will find it difficult to compete with other carriers who claim that they can offer the same service levels. In a recent survey of executives of trucking companies conducted in the Province of Quebec in 1998, it was found that those who reported the best financial performance, as measured by the return on investment, were motor carriers who concentrated their efforts in a niche and those who diversified the range of services offered (Roy and Bigras, 2000). It was also interesting to note that all respondents believed that their company was pursuing a differentiation strategy by providing superior service levels to their customers!

3. A description of motor carrier operations and planning levels

This section first gives a description of the existing motor carrier operations, and discusses the planning decisions associated with different types of carrier operations.

3.1 Motor carrier operations

Intercity trucking companies generally specialize in truckload (TL) freight shipment, less-than-truckload (LTL) freight shipment or parcel shipment.

In TL freight transportation, shipments generally fill an entire trailer. Therefore, they do not require service routing decisions, sorting or handling. Consequently, TL trucking companies do not need to invest in terminals and in specialized city pick-up and delivery equipment. But, due to customer demand, TL carriers are also performing multiple pick-ups with the same destination or one pick-up with multiple deliveries or drops.

LTL carriers typically haul individual shipments weighing between 100 and 10,000 pounds - 50 to 4,500 kg. Since trailers hold 30,000 to 50,000 pounds depending on the density of freight, carriers must consolidate many shipments to make economic use of their vehicles. They have established a large number of terminals to sort freight, both end-of-line terminals at points of origin and destination, and break-bulk terminals, which consolidate freight. End-of-line terminals maintain fleets of small trucks and trailers for handling pick-ups and deliveries in the city. In other countries, like Canada for example, the same terminal usually performs both functions.

Figure 4.1 shows the typical flow of shipments for LTL carriers (Roy and Crainic, 1992). The cycle starts with a demand for a pick-up. These are seldom known in advance and vary from day to day. Demand also varies by season and type of freight. For example, early spring and fall are peak periods in the garment industry. The carrier generally collects shipments in the afternoon and delivers them to the origin terminal, where they are unloaded and verified against the information appearing on the shippers' documentation (bills of lading): weight, dimensions, number of pieces, type of freight, and so forth. The carrier determines tariffs and prepares a waybill. The waybill accompanies the freight and is used to verify it in subsequent handling operations. Next, freight is sorted according to its immediate destination and loaded into line-haul trailers or it is simply moved to the nearest break-bulk terminal if the origin terminal does not sort shipments. Freight addressed to other points in the origin city is transferred to loading zones for local delivery (normally the next morning).

Trailers cannot always be dispatched economically for each destination to which freight is addressed. Such freight is consolidated into trailers going to intermediate terminals where it is loaded with other traffic going to its final destination. In fact, freight may once again be unloaded, sorted, and reloaded at the transfer terminal. Sometimes such freight can be kept in the nose of the trailer and more freight added to it, which reduces the handling costs. Low-volume long-distance shipments may pass through several such intermediate terminals.

At the destination terminal, freight is unloaded, verified, sorted, coded, and moved to loading areas for local delivery to consignees. Line-haul movements occur mostly during the night, so that freight can be delivered in the morning.

Carriers that transport parcels and small shipments such as UPS have many characteristics in common with LTL carriers. They must also

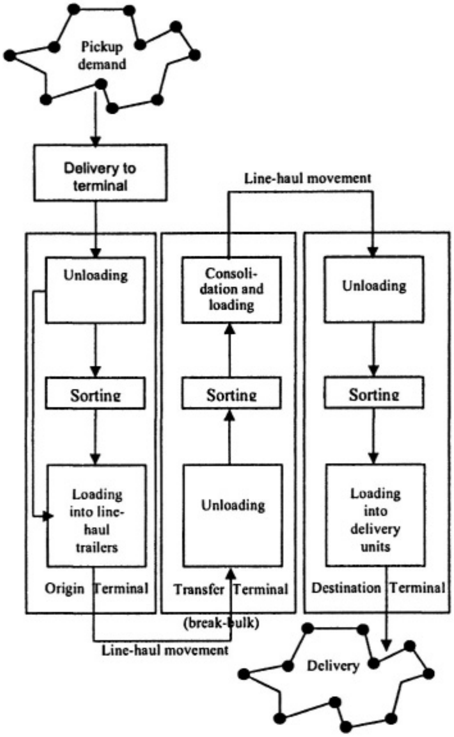


Figure 4.1. Typical flow of LTL shipments.

consolidate shipments and their terminal network is similar to that of an LTL carrier. The remainder of this section will focus on the planning of both TL and LTL operations.

3.2 Levels of planning

The problems motor carriers face in managing their operations can be classified as strategic, tactical, operational or real-time. Several operations research models have been proposed in the scientific literature in order to solve these planning problems. These models have been reviewed in Delorme et al. (1988), Crainic and Laporte (1997), Golden and Assad (1988) and Crainic and Laporte (1998). Table 4.1 provides a sample of the typical problems encountered in trucking companies according to various planning levels.

Strategic planning usually concerns a large part of the organization, has major financial impact, and may have long-term effects. In the motor carrier industry, they typically concern the design of the transportation system: (i) the type and mix of transportation services offered (parcels, LTL or TL services); (ii) the territory coverage and network configuration, including terminal location; and (iii) the service policy, what service levels are offered to customers in terms of both speed and reliability.

Such decisions help determine the motor carrier's strategic position in the market. They should be revised periodically to respond to changes in the environment. The choices carriers make at the strategic level constrain the decision variables at the tactical and operational levels. Facility location is one of the most important strategic decisions and it has thus received a lot of attention (Ballou and Masters, 1993; Miller, 1993).

Tactical planning concerns short- or medium- range activities, such as: (i) equipment acquisition or replacement; (ii) capacity adjustments in response to demand forecasts and the firm's financial situation; and (iii) service network design, freight consolidation, and routing decisions (load plan).

The load plan consists of specifying how freight should be routed over the network. It should consider the following: (i) service network design, the selection of routes on which carrier services will be offered, (ii) freight routing, the sequence of services and consolidation terminals to be used to move freight, and (iii) the movement of empty vehicles, or how to balance the network. Two basic approaches have been proposed in the literature in order to solve the load planning problem: APOLLO (Advanced Planner Of LTL Operations) in Powell and Sheffi (1989) and NETPLAN (NETwork PLANning) in Roy and Delorme (1989). These optimization models have been experimented successfully with actual data from large Canadian and U.S. based LTL trucking companies.

In the motor carrier industry, tactical planning may be performed over a time horizon of a few years when replacing vehicles or purchasing freight-handling equipment, for example. Tactical planning also involves making seasonal adjustments to pick-up and delivery zones, work scheduling and assignments of trucks to the terminal doors in order to adapt to fluctuations in demand over a period of a few months.

Operational planning decisions concern very short-term day-to-day operations. At the operational level, carriers plan current and next day activities. First-line supervisors often make the decisions. Starting with transportation plans formulated at the tactical level, motor carriers assign drivers and equipment and update transportation schedules in response to short term forecasts and daily variations in demand and in equipment and labor availability. The day-to-day planning of delivery routes is one of these problems that has received a lot of attention in the scientific literature under the general designation of vehicle routing problems (Laporte and Osman, 1995).

The hierarchical relationship of these planning levels makes a single model impractical; different formulations are needed to address specific planning levels or problems. At each planning level, the various decisions relate horizontally. Strategies developed for one problem interact with those established for other problems and require that decisions be made globally, network-wide, in an integrated manner (Crainic and Roy, 1988). Finally, in real-time management, carriers combine operational planning and control to maintain the transportation system under equilibrium. This topic will be covered in the next section.

Table 4.1. A sample of typical planning problems in trucking companies.

Planning Levels	**Pickup & Delivery**	**Terminal Activities**	**Long Distance**
Strategic	- Outsourcing - Fleet selection	- Terminal location - Terminal design	- Fleet mix - Outsourcing
Tactical	- P&D zone design - Door assignment	- Load planning - Work scheduling	- Seasonal transportation plan
Operational	- Delivery routes - Truck loading	- Daily adjustments to plans and schedules to plans and schedules	- Vehicle routing - Weekend dispatch
Real-time	- Pick-up assignments - Door assignment	- Hourly adjustments to plans and schedules to plans and schedule	- Real-time dispatch

4. Real-time decision tools

In this section we first describe the issues in real-time planning of less-than-truckload and truckload operations, and discuss what type of decision tools are needed in practice.

4.1 Less-than-truckload operations

Figure 4.2 proposes a description of the real-time planning environment of a LTL carrier. Customer requests for pick-ups are forwarded to the motor carrier, either by phone, fax, EDI or Internet, and the information is registered in the carrier's information system. Pick-up trucks are usually already on the road, in their respective zones, and the customer requests are dispatched to them according to several criteria: space available, type of equipment required, estimated time of arrival at the customer location, distance from the pick-up point, etc. Real-time optimization tools are not extensively used in that context; most motor carriers generally rely on a human dispatcher to make the appropriate decisions on the basis of experience and, sometimes, with the assistance of information systems that display the location of trucks and loads to be picked-up on a map. Although some commercial software packages are becoming available to assist in that matter, very little work has been published thus far on this subject. One of the few papers found in the literature discusses the use of parallel computer processing and a tabu search heuristic to tackle the problem in the context of courier services (Gendreau et al., 1999a). For the LTL application, we refer the reader to Powell, 1988. For a more general review of relevant literature, see also Gendreau and Potvin (1998) and Seguin et al. (1997).

In the more advanced companies, as the loads are picked-up during the day, the relevant information is registered in the on-board computers and forwarded to the terminal where it is fed to the carrier's information system. Although the technology for such real-time information management is available, very few motor carriers are so equipped. Most carriers still have to wait until the pick-up trucks arrive at the terminal at the end of the day in order to find out what was actually picked-up during the day. It is then often too late to perform any kind of optimization at the dock for freight handling and sorting.

To have the information available in real-time is fine but it is even better to use it wisely. To do so, optimization tools can be used daily in order to solve the following problems: (i) assigning pick-up trucks and highway trailers to terminal doors in order to minimize the cost of moving freight across the dock (D' Avignon and Montreuil, 1991); (ii) scheduling the work of freight handling employees (especially part-time workers) according to the daily workload patterns (Nobert and Roy, 1998); and (iii) revising the load plan, i.e., deciding on how freight will be sorted and consolidated at each terminal depending on the actual volumes, per origin-destination pair, picked-up during the day (Roy and Crainic, 1992). Some of these models were initially designed to solve the same problems at the tactical level based on demand forecasts. With the major improvements experienced in computer technology in the last decade, the same models can now be solved very rapidly, in an operational environment, in order to provide daily optimal solutions to motor carrier managers. Using up-to-date information on actual demand, with very short advanced notice, motor carriers can use these static models to assist them in adjusting daily plans in a quasi real-time environment. But, once again, such benefits will be obtained in as much as motor carriers make full use of the information gathered at the pick-up points.

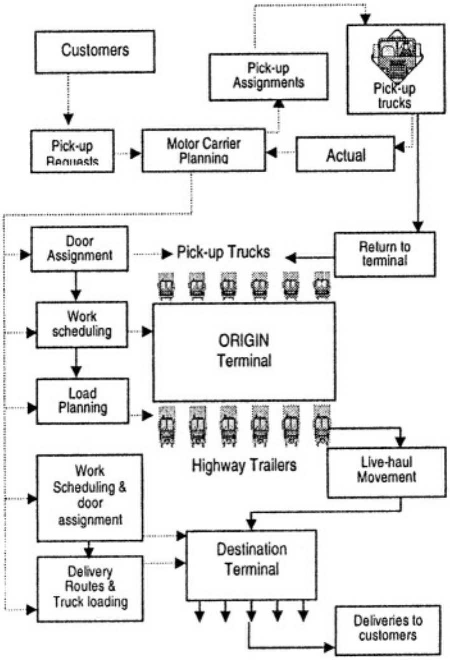


Figure 4.2. Real-time planning of LTL operations.

At the destination terminal, with advanced knowledge of the content of highway trailers, door assignment of trucks and trailers can be adjusted on a daily basis, delivery routes can be redesigned and delivery trucks loaded accordingly. The routing of delivery trucks is in fact an operational problem that has received a lot of attention in the literature as mentioned earlier. But, increasingly, in order to satisfy customer requirements for just-in-time deliveries, motor carriers have to comply with time windows within which deliveries have to be made, or worse, precise appointments that have to be met. Therefore, even though actual demand is known ahead of time, the design of delivery routes in an urban environment with backhauls (pick-ups) and time windows is a difficult problem to solve, but some recent heuristic solutions can be found in Duhamel et al. (1997) and in Thangiah et al. (1996). With the increased emphasis on cross-docking, advanced information on actual freight to be handled will allow motor carriers to improve customer service and reduce costs.

4.2 Truckload operations

Real-time planning of truckload operations is centered around the dispatch of highway tractor-trailer assemblies, as illustrated in Figure 4.3. Customer requests for pick-ups and deliveries are transmitted to the motor carrier, either by phone, fax, EDI or the Internet, and the information is entered in the carrier's information system. In North America, satellite communication has become the standard in the long distance trucking industry. Satellite positioning also allows the carriers to provide up-to-date information to customers with respect to the exact location of their shipments. For example, one of the major truckload companies in the U.S., Schneider National, equipped its fleet with on-board computers and satellite communication devices. Its use of technology has reportedly resulted in cost reductions on the order of 24 percent and increased on-time performance from below 90 to about 99 percent. Using satellite communication, Schneider can track the location of each vehicle anywhere in the U.S. and dispatch it on a real-time basis to satisfy customer needs and adapt to unforeseen changes. This tracking ability has enabled the company to reduce empty mileage by 25 percent. The use of on-board computers also enables Schneider to monitor vehicle speed and drivers' working hours in order to comply with laws and regulations. As a result, technology has helped reduce the company's accident rate by 35 percent since 1987 (Cohen, 1995).

The dynamic allocation of customer requests to vehicles has been studied in Powell (1988) and Powell et al. (1992). The stochastic aspect of the problem has also been considered in Powell et al., 1995 and in Gendreau et al. (1996). These approaches are exploratory and there are still difficulties in solving large-scale real life problems rapidly. The

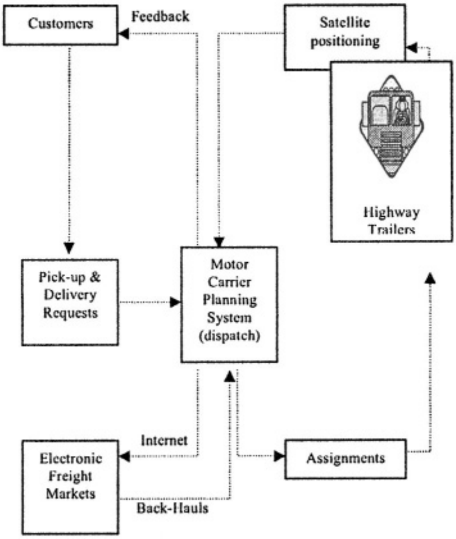

Customers
Feedback
Satellite positioning
Highway Trailers
Pick-up & Delivery Requests
Motor Carrier Planning System (dispatch)
Internet
Electronic Freight Markets
Back-Hauls
Assignments

Figure 4.3. Real-time planning of TL operations.

most common approach in practice is therefore to solve sequences of static assignment problems according to Powell et al. (2002). Savelsbergh and Sol (1998) have reported on the use of such a planning tool, called DRIVE, which has been successfully tested with real-life data in Europe. Alternative dispatch methods to random over the road driver assignments, including the use of zones, key lanes and hub-and-spokes designs, have also been investigated in order to improve drivers' satisfaction and thereby reduce turnover rates at J.B. Hunt (Taylor et al., 1999). But despite the reported benefits of implementing real-time dispatch systems (Regan et al., 1995; Regan et al., 1996) and recent advances in optimization methods and information technology, there is still a relatively small number (a few dozen) of TL motor carriers using these systems today (Powell et al., 2002).

Assigning loads to drivers is a relatively simple optimization problem assuming the data is available, complete and accurate. This author has experienced difficulties in implementing such a decision tool at a major Canadian TL motor carrier mostly because data entry by dispatchers and customer service representatives was deficient. Powell et al. (2002) report on similar implementation problems due to a lack of information made available to the model, even though significant potential benefits can be achieved by using the system properly.

Real-time decision tools are also rapidly becoming available on the Internet to assist motor carriers in planning their work in a real-time environment. *E-Dispatch* and *E-Scheduling* are now very much part of the transportation vocabulary. Also available through the Internet is the possibility for providers of freight services to make electronic bids for trips offered through electronic freight markets. These are of particular interest for back-hauls, thus improving the efficiency of carriers by minimizing waiting time and empty mileage. These one-way trips or freight movements are usually sold individually. This process is thus equivalent to a conventional sealed-bid auction in which a group of bidders (trucking companies) submit their price for a given freight movement in a sealed envelope and then the lowest quote is selected by the shipper. More recently, software has been developed to allow shippers to offer all available freight movements simultaneously and carriers to bid on a combination of freight movements at one time; this process being referred to as combinatorial auction (Song and Regan, 2002). The benefits are plain to see: when a carrier is offered several freight movements (instead of a single one), it can increase the chances of generating more efficient routes when adding these movements to its actual planned trips (Caplice, 1996).

5. Concluding remarks and future research directions

This chapter has reviewed some of the most important recent trends in supply chain management and examined their impact on motor carriers. It was found that the increased used of time-sensitive strategies such as JIT, QR, and ECR, along with the development of electronic commerce and the outsourcing of logistical services, has and will continue to have a major impact on motor carriers. Indeed, carriers will have to adapt by forming partnerships with customers, providing logistical services, investing in new information and communication technologies, and using real-time decision tools.

Next, the typical operations of a motor carrier were described, and some of the more "traditional" tactical and operational planning tools were quickly referenced. The real-time planning environment of both LTL and TL carriers was then described. It was found that several of the tools developed for the tactical planning level could be adapted and used to address some of the operational and even real-time (short advance notice) issues such as assigning trucks and trailers to terminal doors, scheduling part-time freight handling employees and revising load plans. It was also found that a very small number of motor carriers actually make use of the information that can be gathered at the pick-up points through on-board computers and real-time information systems. We believe that using this information wisely in conjunction with real-time and operational decision tools will significantly improve the performance of motor carriers, just like it did in the case of courier companies and just like new logistical practices rest on the use of point-of-sale information.

Future research efforts are, however, required to better understand why so many motor carriers are still reluctant to reap the benefits associated with the implementation of new technologies, including the decision support tools that have proven to generate significant savings to the industry. The recent paper by Powell et al. (2002) sheds some light on some of the managerial, informational and behavioral issues related to the implementation of such decision support tools in the motor carrier industry. Further research is also needed on the role and importance of freight carriers in the implementation of modern supply chain practices such as CRP, VMI and CPFR. In particular, it would be interesting to determine to what extent carriers should be integrated into the replenishment planning process and how their role could evolve from simple transportation to logistics services providers. Finally, more work is required on the development of user-friendly decision support tools in a real-time environment and on the impact of recent advances in combinatorial auctions on the trucking industry.

References

Ballou, R. and Masters, J.M. (1993). Commercial software for locating warehouses and other facilities. *Journal of Business Logistics,* 14(2): 71–107.

Bowersox, D.J., Closs, D.J., and Cooper, M.B. (2002). *Supply Chain Logistics Management.* McGraw-Hill, New York, NY.

Bowersox, D.J. and Taylor, J.C. (2001). World trade to become more intra-regional. In *Proceedings of the 22nd ICHCA Biannual Conference, 2001 A Transportation Odyssey, Toronto,* pages 2–6.

Caplice, C. (1996). *An Optimization Based Bidding Process: A New Frameworkfor Shipper-Carrier Relationship.* PhD thesis, Massachusetts Institute of Technology, Boston, Massachusetts.

Chow, G. and Gritta, R. (2002). The north american logistics service industry. Proceedings of the Fourth International Meeting for Research in Logistics, Lisbon, Portugal. CD Rom.

Closs, D.J., Goldsby, T.J., and Clinton, S.R. (1997). Technology influences on world class-logistics capability. *International Journal of Physical Distribution and Logistics Management,* 27(1): 4–17.

Cohen, W. (1995). Taking to the highway. *US News and World Report,* pages 84–87.

Crainic, T.G. and Laporte, G. (1997). Planning models for freight transportation. *European Journal of Operational Research,* 97(3): 409–438.

Crainic, T.G. and Laporte, G., editors (1998). *Fleet Management and Logistics.* Kluwer Academic Publishers, Boston, Massachusetts.

Crainic, T.G. and Roy, J. (1988). O.R. tools for tactical freight transportation planning. *European Journal of Operational Research,* 33(3): 290–297.

D'Avignon, G. and Montreuil, B. (1991). La minimisation de l'effort de transbordement des marchandises d'un terminus routier. In *Proceedings of the Administrative Sciences Association of Canada, Niagara Falls, Ontario.*

Delorme, L., Roy, J., and Rousseau, J.M. (1988). Motor carriers operations planning models: A state of the art. In Bianco, L. and Labella, A., editors, *Freight Planning and Logistics,* pages 510–545. Springer-Verlag, Boston, Massachusetts.

Dismukes, T. and Godin, P. (1997). How information technology & electronic commerce will shape logistics in the 21st century. *Logistics Quarterly,* 3(3): 18–20.

Duhamel, C., Potvin, J.Y., and Rousseau, J.M. (1997). A tabu search heuristic for the vehicle routing problem with backhauls and time windows. *Transportation Science,* 31(1): 64–59.

Factor, R. and Kilpatrick, J. (1998). Logistics in canada: A survey. *Materials Management and Distribution,* pages 16–27.

Gendreau, M., Guertin, F., Potvin, J.Y., and Taillard, E. (1999a). Parallel tabu search for real-time vehicle routing and dispatching. *Transportation Science,* 33(4): 381–390.

Gendreau, M., Laporte, G., and Séguin, R. (1999b). Stochastic vehicle routing. *European Journal of Operational Research,* 88: 3–22.

Gendreau, M. and Potvin, J.Y. (1998). Dynamic vehicle routing and dispatching. In Crainic, T.G. and Laporte, G., editors, *Fleet Management and Logistics,* pages 115–126. Kluwer Academic Publishers, Boston, Massachusetts.

Golden, B.L. and Assad, A.A., editors (1988). *Vehicle Routing: Methods and Studies.* North-Holland, Amsterdam.

Golob, T.F. and Regan, A.C. (2001). Impacts of information technology on personal travel and commercial vehicle operations: Research challenges and opportunities. *Transportation Research Part C: Emerging Technologies,* 9: 87–121.

Golob, T.F. and Regan, A.C. (2002). Trucking industry adoption of information technology: A multivariate discrete choice model. *Transportation Research Part C: Emerging Technologies,* 10: 205–228.

Gopal, C. and Fox, M.L. (1996). Electronic commerce and supply chain reengineering. In *Proceedings of the Council of Logistics Management Annual Conference, Orlando, Florida.*

Laporte, G. and Osman, I.H. (1995). Routing problems: A bibliography. *Annals of Operations Research,* 61: 227–262.

Lieb, R.C. and Peluso, L.A. (1999). The use of third party logistics services by large american manufacturers: The 1999 survey. In *Proceedings of the 1999 Council of Logistics Management Annual Conference, Toronto, Canada,* pages 159–171.

Masters, J.M., LaLonde, B.J., and Williams, L.R. (1991). The effect of new information technology on the practice of trafic management. *The International Journal of Logistics Management,* 2(2): 13–21.

Miller, T. (1993). Learning about facility location models. *Distribution,* 92(5): 47–50.

Nobert, Y. and Roy, J. (1998). Freight handling personnel scheduling at air cargo terminals. *Transportation Science,* 32(2): 295–301.

Powell, W.B. (1988). A comparative review of alternative algorithms for the dynamic vehicle allocation problem. In Golden, B.L. and Assad, A.A., editors, *Vehicle Routing: Methods and Studies,* pages 297–373. North-Holland, Amsterdam.

Powell, W.B., Jaillet, P., and Odoni, A. (1995). Stochastic and dynamic networks and routing. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Handbooks in Operations Research and Management Science - Network Routing,* pages 141–295. North-Holland, Amsterdam.

Powell, W.B., Marar, A., Gelfang, J., and Bowers, S. (2002). Implementing real-time optimization models: A case application from the motor carrier industry. *Operations Research,* 50(4): 571–581.

Powell, W.B. and Sheffi, Y. (1989). Design and implementation of an interactive optimization system for network design in the motor carrier industry. *Operations Research,* 37(1): 12–29.

Powell, W.B., Sheffi, Y., Nickerson, K.S., Butterbaugh, K., and Atherton, S. (1992). Maximizing profits for North American Van Lines' Truckload Division: A new framework for pricing and operations. *Interfaces,* 18(1): 21–41.

Raney, M.A. and Walter, C.K. (1992). Electronic data interchange: The warehouse and supplier interface. *International Journal of Physical Distribution and Logistics Management,* 22(8): 21–26.

Ratliff, H.D. (1995). Logistics management: integrate your way to an approved bottom line. *IIE Solutions,* 27(10): 31–33.

Regan, A.C., Mahmassani, H.S., and Jaillet, P. (1995). Improving efficiency of commercial vehicle operations using real-time information: Potential uses and assignment strategies. *Transportation Research Record,* 1493: 188–198.

Regan, A.C., Mahmassani, H.S., and Jaillet, P. (1996). Dynamic decision making for commercial fleet operations using real-time information. *Transportation Research Record,* 1537: 91–97.

Rogers, D.S., Daugherty, P.J., and Stank, P.T. (1993). Enhancing service responsiveness: The strategic potential of EDI. *Logistics Information Management,* 6(3): 27–32.

Roy, J. and Bigras, Y. (2000). Competitiveness factors and strategies of motor carriers in quebec. Working paper, Centre de recherche en gestion, Université du Québec à Montréal, Montrèal, Canada.

Roy, J. and Crainic, T.G. (1992). Improving intercity freight routing with a tactical planning model. *Interfaces,* 22(3): 32–44.

Roy, J. and Delorme, L. (1989). NETPLAN: A network optimization model for tactical planning in the less-than-truckload motor carrier industry. *INFOR,* 27(1): 22–35.

Savelsbergh, M. and Sol, M. (1998). DRIVE: Dynamic routing of independent vehicles. *Operations Research,* 46(4): 474–490.

Séguin, R., Potvin, J.Y., Gendreau, M., Crainic, T.G., and Marcotte, P. (1998). Real-time decision problems: An operational research perspective. *Journal of the Operational Research Society,* 48: 162–174.

Simchi-Levi, D., Kaminsky, P., and Simchi-Levi, E., editors (2003). *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies.* McGraw-Hill, New York, New York.

Sink, H.L. and Langley, Jr., C.J. (1997). A managerial framework for the acquisition of third party logistics services. *Journal of Business Logistics,* 18(2): 163–189.

Song, J. and Regan, A.C. (2002). Combinatorial auctions for transportation service procurement: The carrier perspective. Working Paper UCI-ITS-LI-WP-02-8, Institute of Transportation Studies, University of California, Irvine, California.

Sunstrum, A.L. and Howard, F.H. (1996). Strategic alliances enabled by electronic commerce. In *Proceedings of the 31st Annual Meeting of the Canadian Transportation Research Forum, Winnipeg, Canada,* pages 252–265.

Taylor, G.D., Meinert, T.S., Killian, R.C., and Whicker, G.L. (1999). Development and analysis of alternative dispatching methods in truckload trucking. *Transportation Research Part E: Logistics and Transportation Review,* 35(3): 191–205.

Thangiah, S.R., Potvin, J.Y., and Sun, T. (1996). Heuristic approaches to vehicle routing with backhauls and time windows. *International Journal on Computers and Operations Research,* 113: 1043–1057.

Udo, G.J. and Pickett, G.C. (1994). EDI conversion mandate: The big problem for small businesses. *Industrial Management,* 36(2): 6–9.

Walton, L.W. (1994). Electronic data interchange (EDI): A study of its usage and adoption within marketing and logistics channels. *Transportation Journal,* 34(2): 37–45.

Winters, J. (1999). The effective implementation of co-managed inventory. In Fernie, J. and Sparks, L., editors, *Logistics and Retail Management: Insights Into Current Practice and Trends From Leading Experts,* pages 141–153. Kogan, London, United Kingdom.

Chapter 5

MANAGING THE SUPPLY-SIDE RISKS IN SUPPLY CHAINS: TAXONOMIES, PROCESSES, AND EXAMPLES OF DECISION-MAKING MODELING

Amy Z. Zeng
Department of Management
Worcester Polytechnic Institute
Worcester, Massachusetts 01609
azeng@wpi.edu

Paul D. Berger
School of Management
Marketing Department
Boston University
Boston, Massachusetts 02215
pdberger@bu.edu

Arthur Gerstenfeld
Department of Management
Worcester Polytechnic Institute
Worcester, Massachusetts 01609
ag@wpi.edu

Abstract Today's supply chains are becoming not only more efficient with the aid of advanced information technologies, but also riskier due to the tight interconnectedness of numerous chain links that are prone to breakdowns, disruptions or disasters. Although many studies focusing on business risks in various contexts have been presented in the literature over the years, research effort devoted to understanding the risks associated with suppliers and the supply market has been limited, especially from a quantitative point of view. In this chapter, we first, through extensive literature view, present a taxonomy of supply-side risks, a four-step supply risk management process, and a list of techniques that help accomplish each step. Then we propose two optimization-based decision tree models that effectively formulate two decision-making situations in which the questions of how many suppliers should be used and whether to use standby suppliers are addressed. Future research directions are also suggested.

1. Introduction

Triggered by the recognition that it is imperative to focus on core competencies to remain competitive in the market, organizations are forming seamless supply chains to manage the flows of material, information, and capital from the earliest supplier of raw materials to the ultimate customer. This integrative philosophy, termed as supply chain management, has led numerous companies to improved performance, thereby gaining widespread popularity and adoption as a top priority in nearly every enterprise.

It is straightforward to see that the central theme of supply chain management is the integration of all activities spanning from the physical supply to physical distribution through improved supply chain relationships. Along each supply chain, numerous links exist, embracing material suppliers, production facilities, distribution services, warehouses, service centers, and customers into a cohesive entity. These chain partners coordinate and collaborate with their decisions and activities, thus creating sustainable competitiveness. However, as the efficiency of supply chains increases, so does the likelihood of a breakdown at any point on the chain continuum, due in part to the heavy inter-dependence of the chain partners, heightened competition, lengthened supply and distribution lines, and more knowledgeable customers. Therefore, to reduce the likelihood of a breakdown, or plan for unavoidable disruptions, a company must carefully deal with its supply chain risks and uncertainties. Recent terrorist attacks especially signify the necessity and importance of supply chain risk management.

Since risk can be defined as the danger that events or decisions will obstruct a company's achievement of its objectives (Zsidisin, 2001), and a disruption can be viewed as a situation during the operations' execution in which the deviation from plan is sufficiently large that the plan has to be changed substantially (Clausen et al., 2001), we treat these two terms as interchangeable. Risk management is undoubtedly a broad subject and has been around for a few decades. Numerous discussions and studies over the years have examined business risks and disruptions in the contexts of financial planning, new-product development, and demand changes. However, as pointed out by Zsidisin (2001), there has been a limited amount of attention given to understanding the risks associated with suppliers and the supply networks. As the degree of reliance on the suppliers is accelerating and the physical supply cost can easily account for 30% of the total cost of goods sold (Ballou, 1999), higher risks can be posed for purchasing managers, the entire organization, and even the whole supply chain (Smeltzer and Siferd, 1998). Therefore, efficient management of the supply-side risks contributes significantly to ultimate supply chain success.

Thus, the objective of this chapter is fourfold. First of all, we summarize the key risk factors pertinent to supply networks, based on our extensive literature review. Secondly, a four-step management process dealing especially with the supply risks is described in detail, along with a set of techniques that can be used to accomplish each step. Thirdly, we present two decision-tree based optimization models to show how decisions on the number of suppliers and the use of standby suppliers can be made. Finally, a number of research directions for studying the supply-side risks are suggested.

2. Supply-side Risk Taxonomy

Previous studies have summarized the major enterprise risk factors that are inherent with a corporation of any kind. For example, Grey and Shi (2001) have classified those enterprise risks into two categories: core business risk, which consist of value chain risk and operational risk, and non-core business risk, which comprises event risk and recurring risk. With regard to the supply-side risk in particular, a number of ways of categorization suggested in existing studies can also be found, four of which are briefly described here. Zsidisin and Ellram (1999) rely on a case study to group the supply risk factors into incoming materials, source process capability, volume capacity, technologically capable, and supplier lead time. Another study by Zsidisin et al. (2000) summarizes the key supply risks into the following list: business risk, supplier capacity constraints, quality, production technological changes, product design changes, and disasters. Michalski (2000) presents how Microsoft identifies and organizes risks into four families: infrastructure, business controls, business value, and relationship. A fourth example given by Johnson (2001) simply considers the entire supply chain risks as two groups: supply risks and demand risks, where the supply risks come from the manufacturing capacity, logistics capacity, currency fluctuations, and supply disruptions from political issues.

In this chapter, we attempt to combine various ways of supply risk classification found in the literature and present our taxonomy in Figure

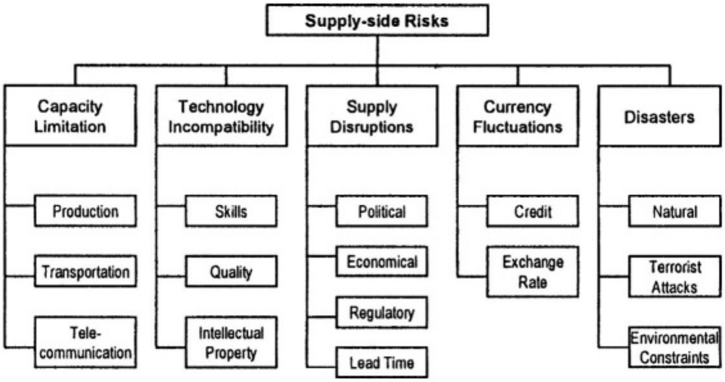


Figure 5.1. Supply-side Risk Taxonomy.

5.1. Our intention is not to provide an exhaustive list of supply risk factors; rather to present the supply risk families that may be experienced by various industry sectors. Specifically, we classify the supply risks into five families: capacity related, technology related, supply related, currency related, and disasters related, each of which consisting of a number of risk elements. These risk families not only consider the capacity limitation and technology gaps existing in potential suppliers, but also take into account the locations and external impacts of the suppliers. For instance, supply disruptions resulting from political, economical and regulatory influences, and long lead time are especially prominent in oversea suppliers (Dornier et al., 1998; Souter, 2000). And currency fluctuations pertinent to credit and exchange rate are especially thorny issues when employing international suppliers residing in developing countries (Dornier et al., 1998). Additionally, we also recognize that natural disasters as well as human planned attacks, such as the terrorist attack of 09/11/01, and environmental constraints, can all be catastrophic for the functionality of the supply networks. We hope that this supply-side risk taxonomy will provide a useful guideline for companies to identify the potential risks associated with their suppliers or supply markets.

3. Supply Risk Management Process

Understanding the potential risk factors residing in the supply networks is not sufficient, and to mitigate and monitor the impact of the risks, an efficient risk management process must be in place to support an organization's supply chain strategies. This management process for

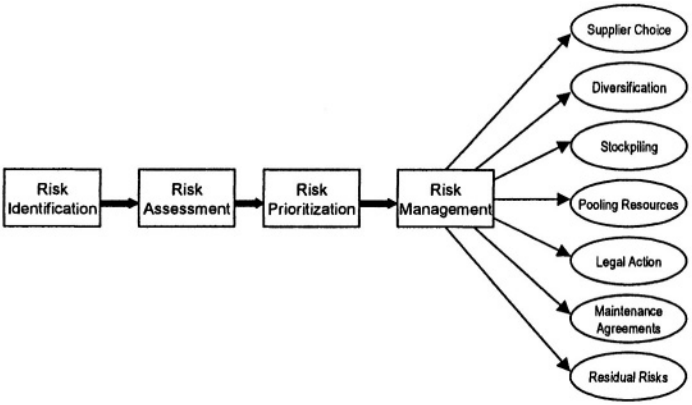
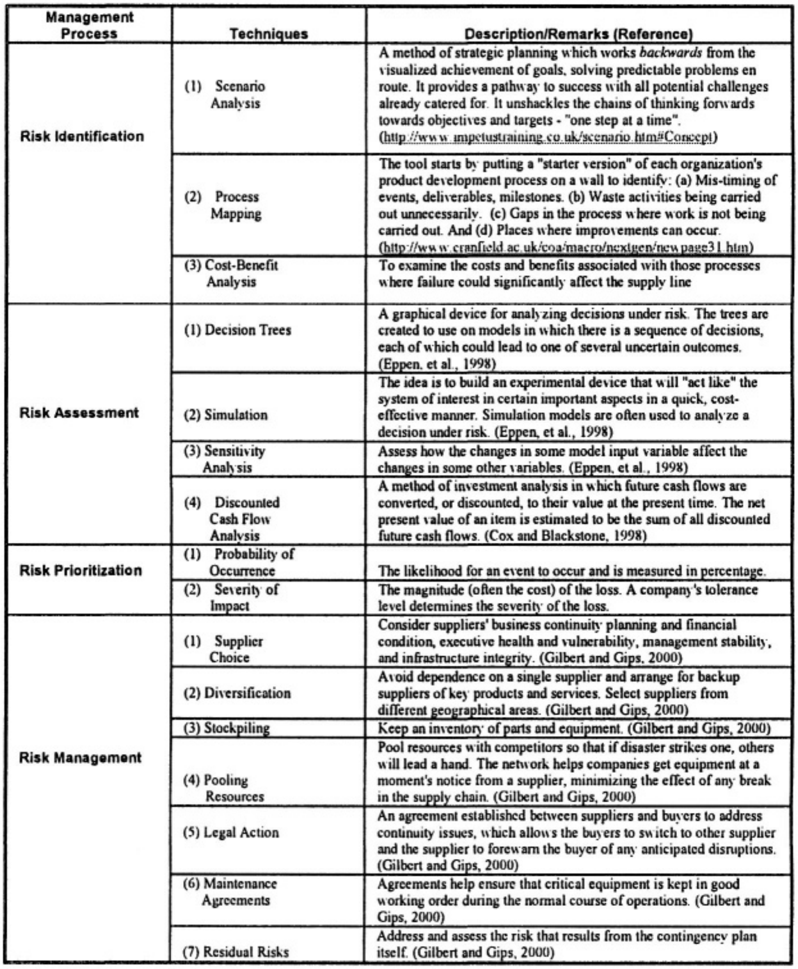


Figure 5.2. Supply Risk Management Process.

coping with the supply risks does not need to be "rocket science"; rather it simply requires an implementation of the traditional risk management process (Souter, 2000). A number of examples of the management process have been proposed in the literature (e.g., Gilbert and Gips, 2000; Souter, 2000; Grey and Shi, 2001), possessing substantial similarities. We integrate these example processes and present a four-step procedure in Figure 5.2, which shows that managing the supply risks starts with risk identification, continues through risk assessment and risk prioritization, and ends with risk management strategy formulation. To deal with the risks associated with the suppliers, Gilbert and Gips (2000) have summarized an excellent list of methods for reducing the risk impacts; the list is included in Figure 5.2. To complete each step in the process, a set of quantitative techniques is often required to assist the analysis, and a balance between quantitative and qualitative analyses is also desired. We list some of the frequently used techniques for each of the four steps with brief descriptions in Table 1. It is seen that risks can be identified through scenario analysis, process mapping, and cost-benefit analysis, and assessed mathematically not only by some management science based techniques such as decision trees, simulation, and sensitivity analysis, but also by accounting methods such as discounted cash flow analysis. The objective of risk prioritization is to rank the risks based on their probability of occurrence (high versus low) and severity of impacts (high versus low). The two dimensions can result in a two-by-two matrix, with each quadrant indicating a management direction for companies to take. For example, Grey and Shi (2001) advise that for risks with low severity and low likelihood of occurrence, it is sufficient to monitor pe-

Table 1: Techniques Associated with Each Supply Risk Management Process



Management Process	Techniques	Description/Remarks (Reference)
Risk Identification	(1) Scenario Analysis	A method of strategic planning which works *backwards* from the visualized achievement of goals, solving predictable problems en route. It provides a pathway to success with all potential challenges already catered for. It unshackles the chains of thinking forwards towards objectives and targets - "one step at a time". (http://www.impetustraining.co.uk/scenario.htm#Concept)
	(2) Process Mapping	The tool starts by putting a "starter version" of each organization's product development process on a wall to identify: (a) Mis-timing of events, deliverables, milestones. (b) Waste activities being carried out unnecessarily. (c) Gaps in the process where work is not being carried out. And (d) Places where improvements can occur. (http://www.cranfield.ac.uk/coa/macro/nextgen/newpage31.htm)
	(3) Cost-Benefit Analysis	To examine the costs and benefits associated with those processes where failure could significantly affect the supply line
Risk Assessment	(1) Decision Trees	A graphical device for analyzing decisions under risk. The trees are created to use on models in which there is a sequence of decisions, each of which could lead to one of several uncertain outcomes. (Eppen, et al., 1998)
	(2) Simulation	The idea is to build an experimental device that will "act like" the system of interest in certain important aspects in a quick, cost-effective manner. Simulation models are often used to analyze a decision under risk. (Eppen, et al., 1998)
	(3) Sensitivity Analysis	Assess how the changes in some model input variable affect the changes in some other variables. (Eppen, et al., 1998)
	(4) Discounted Cash Flow Analysis	A method of investment analysis in which future cash flows are converted, or discounted, to their value at the present time. The net present value of an item is estimated to be the sum of all discounted future cash flows. (Cox and Blackstone, 1998)
Risk Prioritization	(1) Probability of Occurrence	The likelihood for an event to occur and is measured in percentage.
	(2) Severity of Impact	The magnitude (often the cost) of the loss. A company's tolerance level determines the severity of the loss.
Risk Management	(1) Supplier Choice	Consider suppliers' business continuity planning and financial condition, executive health and vulnerability, management stability, and infrastructure integrity. (Gilbert and Gips, 2000)
	(2) Diversification	Avoid dependence on a single supplier and arrange for backup suppliers of key products and services. Select suppliers from different geographical areas. (Gilbert and Gips, 2000)
	(3) Stockpiling	Keep an inventory of parts and equipment. (Gilbert and Gips, 2000)
	(4) Pooling Resources	Pool resources with competitors so that if disaster strikes one, others will lead a hand. The network helps companies get equipment at a moment's notice from a supplier, minimizing the effect of any break in the supply chain. (Gilbert and Gips, 2000)
	(5) Legal Action	An agreement established between suppliers and buyers to address continuity issues, which allows the buyers to switch to other supplier and the supplier to forewarn the buyer of any anticipated disruptions. (Gilbert and Gips, 2000)
	(6) Maintenance Agreements	Agreements help ensure that critical equipment is kept in good working order during the normal course of operations. (Gilbert and Gips, 2000)
	(7) Residual Risks	Address and assess the risk that results from the contingency plan itself. (Gilbert and Gips, 2000)

riodically for change in status; and for risks with high likelihood but low severity, it is important to deploy operational changes and controls to reduce the frequency of occurrence. However, it is those risks with high severity of impact that companies must pay great attention to and make contingency plans for. Once the risks are ranked and prioritized, a company is ready to determine recovery time and choose appropriate strategies for various types of incidents. As mentioned before, Gilbert and Gips (2000) have suggested seven possible methods for managing the risks, which include selection of low-risk suppliers, diversification of suppliers, stockpiling, resource pooling, legal action, maintenance agreement between buying and supply firms, and consideration of the residual effects resulting from the contingency plans employed.

We believe that the four-step supply risk management process presented in Figure 5.2 and the associated techniques listed in Table 1 provided excellent starting points for companies involved in supply chains and concerned with supply risks. The strategic use of the seven methods are discussed extensively in the literature, however, we would like to take one step further in this chapter - we will present optimization-based mathematical models to aid in decision-making with regard to two of the methods, namely, using several suppliers (diversification) and pooling resources (establishment of a supply network).

4. Decision-making Modeling: Two Examples

Previous sections have presented supply-risk taxonomy, risk management process, and important techniques for companies along the supply chain to reduce the likelihood of a breakdown or to mitigate impact of disasters at any point in the supply lines. In this section, we propose two decision-tree models that can help a company to make decisions on the optimal number of suppliers and the optimal use of standby suppliers. It is necessary to point out that the motivations of the studies, related literature review, and detailed numerical examples and sensitivity analysis can be found in papers by Berger et. al. (2002a, 2002b) and Zeng and Berger (2003), and hence are not repeated; only the modeling details are described in this section.

4.1 How Many Suppliers Are Best?

As the preceding section discussed, one of the strategies a company can employ to deal with the supplier risks is to diversify the supply base, that is, to keep multiple sources available for key products or services, which helps not only prevent emergencies, but promotes competitive bidding. While this issue falls well into the vehement debate on single sourcing or multiple sourcing, there has been limited *quantitative* study on selecting the number of sources when certain supplier risks should be taken into consideration. We intend to bridge this gap by using a decision-tree based model to formulate the decision-making situation and then to identify the optimal number of sources a buying firm should have.

4.1.1 The Decision-Tree Approach. To jointly consider the risk and cost associated with using multiple suppliers or single supplier, we model the decision-making process using a decision tree as illustrated in Figure 5.3. The following are elements of the decision tree:

1 The "acts" are the choices of number of suppliers, ranging from i = 1 to n, where n is a decision variable;

2 For each choice, there are two states of nature: *all suppliers are down* and *not all suppliers are down.*

3 When all suppliers are down, the financial loss to the decision-making company (that is, the buying firm) is given by a constant, L; and the cost of operating i suppliers is given by $C(i)$, $i = 1, 2 \ldots n$.

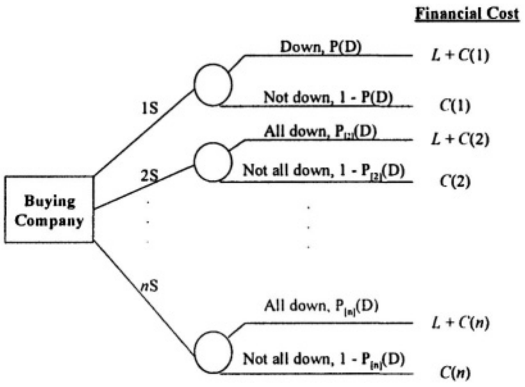
Therefore, according to the basic decision-tree analysis, ETC(n), the expected total cost experienced by the buying firm when n suppliers are used, is

$$\begin{aligned} \mathrm{ETC}(n) &= P_{[n]}(D)[L + C(n)] + [1 - P_{[n]}(D)]C(n) \\ &= C(n) + L \cdot P_{[n]}(D) \end{aligned} \tag{5.1}$$

It is seen that the properties of the expected total cost function in (5.1) depends on the form of two elements: C(n) and $P_{[n]}$(D). Intuitively, the operating cost, C(n), is an increasing function of the number of suppliers employed, whereas, the expected financial loss, $L \cdot P_{[n]}(D)$, decreases with the number of suppliers increases, as more suppliers provide a higher level of reliability; mathematically,

$$P_{[n]}(\mathrm{D}) - P_{[n-1]}(\mathrm{D}) < 0. \tag{5.2}$$

Therefore, a trade-off exists of these two cost components, and the optimal number of suppliers needs to be determined. In what follows, we will examine the characteristics of (5.1) using some assumed functions for the two elements.



Notation:

*n*S:	*n* suppliers, *n* = 1, 2, ..., a decision variable
P(D):	probability of down
$P_{[n]}(D)$:	probability that *n* suppliers are down
L:	financial loss caused by all suppliers being down
C(*n*):	operating cost of *n* suppliers incurred to the buying company

Figure 5.3. A Decision Tree Model for Determining the Optimal Number of Suppliers.

4.1.2 Basic Scenario: Constant Probability and Linear Operating Cost. The major reason for the potential superiority of a buying company having more than one supplier is that if one supplier is "down (D)" [i.e., indeed, is not able to supply the company's needs], another supplier may be able to "pick up the slack" and provide the company's needs. We presume that there are potential "super-events" that can occur affecting *all* suppliers simultaneously. That is, events such as terrorism or a widespread airline action that put *all* suppliers down. The probability of one of these super-events occurring during the supply cycle is denoted by *P*. Of course, we also include a "unique-event" scenario for each supplier; that is, an event uniquely associated with a particular supplier that puts it down during the supply cycle. We designate this probability as U_i for supplier i, $i = 1, 2, \ldots n$. We have not included the situation when there are "semi-super events" – that is, events that affect a subset of all suppliers (but more than *one*), but not all suppliers. We are confident that this "omission" does not materially affect the salient issues to be analyzed and discussed, while it greatly simplifies our exposition. Also, given our formulation, by definition, U_i and U_j are independent for $i \neq j$. And, we reasonably assume that U_i and *P* are independent events. Then, the probability that supplier i is down is $\mathrm{P}_i(\mathrm{D}) = P + (1 - P) \cdot U_i$, and the probability that all n ($n = 1, 2, \dots$) suppliers are down is expressed as

$$P_{[n]}(D) = P + (1 - P) \cdot U_1 \cdot U_2 \cdot \dots \cdot U_n \tag{5.3}$$

Assume for the ease of exposition that the unique probability of down is about the same for each supplier[1], that is, $U_1 = U_2 = \dots = U_n = U$. Furthermore, the buying firm's operating cost with multiple suppliers is a linear function, that is, $\mathrm{C}(n) = a + b(n)$, where $\mathrm{C}(n)$ is a linear function of n. Thus, the general cost function in (5.1) in this basic scenario with linear operating cost and event-based probability is simplified to

$$\mathrm{ETC_L}(n) = (a + b \cdot n) + L \cdot [P + (1 - P) \cdot U^n] \forall \begin{cases} 0 < P,\ U < 1 \\ a,\ b > 0 \end{cases} \tag{5.4}$$

It is straightforward to prove that the cost function in (5.4) is convex and has a minimum. The optimal solution for n in a continuous form, n_{C}^*, is found as follows:

$$n_{\mathrm{C}}^* = \frac{\ln [b/(L \cdot (1 - P) \cdot (-\ln U))]}{\ln U}. \tag{5.5}$$

Hence, the value of n^* as an integer, n_{I}^*, should satisfy

$$\mathrm{ETC}(n_{\mathrm{I}}^*) = \min \{\mathrm{ETC_L} \lfloor n_{\mathrm{C}}^* \rfloor, \mathrm{ETC_L} \lceil n_{\mathrm{C}}^* \rceil\}. \tag{5.6}$$

4.1.3 Scenario Two: Non-linear Operating Cost and Probability.

In this section we examine the properties of the expected cost function in (5.1) with more general operating cost and probability density functions. We first consider the operating cost of doing business with multiple suppliers. It is reasonable to treat each supplier as an additional plant for the buying firm. According to Nahmias (2001), the cost of operating an additional plant experienced in a variety of industries can be represented by a power function as

$$C(y) = ky^{\alpha}, k > 0 \text{ and } 0 < \alpha < 1,$$

[1] We think that this assumption is fairly realistic. Each company is facing various kinds of risks; however, although the "unique-events" causing a supplier to become unavailable may be different for each supplier, the probability of the combined risk effects will likely be similar for all suppliers.

where y is the supplier's supply capacity, k is a constant of proportionality, the exponent α measures the ratio of the incremental to the average cost of a unit supplier capacity, that is, $0 < \alpha < 1$. If the buying firm deals with multiple suppliers, then the total operating cost is given by

$$C(n) = k\left(\sum_{i=1}^{n} y_i\right)^{\alpha}. \tag{5.7}$$

It is seen that the cost in (5.7) is difficult to deal with; thus, we offer an approximation. Assume that the capacity of all available suppliers, y_i, can be ranked in an ascending order with the smallest supply capacity being y_s, and furthermore,

$$y_i = (1+r)y_{i-1}, 0 < r < 1 \text{ and } i = 1, 2, \ldots, n$$

then,

$$\begin{aligned} C(n) &\approx ky_s^{\alpha}\left[1 + (1+r) + (1+r)^2 + \ldots + (1+r)^{n-1}\right]^{\alpha} \\ &= k \cdot (y_s/r)^{\alpha} \cdot [(1+r)^n - 1]^{\alpha} \end{aligned} \tag{5.8}$$

It can be shown (see Zeng and Berger, 2003) that the approximate operating cost in (5.8) is a strictly increasing convex function with respect to the decision variable, n.

Now we consider the formulation of $\mathrm{P}_{[n]}(\mathrm{D})$ in (5.1). In the basic scenario we considered the super and unique events that cause all suppliers to become down to satisfy the buying firm's demand. To obtain closed-form results, one needs to assume that the probability of the unique event causing a specific supplier down is identical for all employed suppliers. In this scenario, instead of looking at those events that bring the suppliers to be down, we assume that the number of unavailable suppliers follows a discrete probability distribution. Since the Poisson distribution has been widely used in many similar scenarios, we choose to use a Poisson with parameter μ to model the probability that there are n (n = 1, 2, ...) suppliers down and unavailable to meet the buying firm's needs due to various reasons. Suppose that it is possible to estimate the average number of unavailable suppliers, μ, based on the firm's past experiences with the suppliers; then, we can write the probability as follows:

$$\phi_{[n]}(D) = \frac{e^{-\mu} \cdot \mu^n}{n!}, n = 1, 2, \ldots \tag{5.9}$$

To meet the requirement displayed in (5.2), it is necessary, as can be shown, that $n > \mu$, meaning that the number of suppliers employed during a supply cycle should be greater than the average number of unavailable suppliers. This requirement is commonsense, and thereby poses no important restrictions; however, this simple requirement immediately indicates that as long as the average number of suppliers unable to meet the buying firm's demand is greater than one, the buying firm should consider using multiple suppliers.

According to Zeng and Berger (2003), the financial loss caused by all suppliers that are unable to satisfy the buying firm's demand, $L(n) = L \cdot \phi_{[n]}(D)$, is a decreasing concave function with maximum occurring at $n = 1$ and minimum existing at $n = \infty$.

Now substituting the approximate operating cost in (5.8) and the Poisson density function in (5.9) to the general expected total cost function in (5.1) gives the following:

$$\mathrm{ETC_p}(n) = k\left(\frac{y_s}{r}\right)^{\alpha}((1+r)^n - 1)^{\alpha} + L\frac{e^{-\mu}\mu^n}{n!}, \forall \begin{cases} k > 0 \\ 0 < r, \alpha < 1 \end{cases} \tag{5.10}$$

The special case of (5.10) occurs at $n = 1$, where $\mathrm{ETC_p}(1) = ky_s + Le^{-\mu}\mu$. Additionally, the possible scenarios when finding the optimal value of n are as follows:

1 If $ky_s^{\alpha} \gg L$, that is, the operating cost is much larger than the financial loss or the total cost function is dominated by the operating cost, then, $n^* = 1$;

2 If $L \gg ky_s^{\alpha}$ and r is very small, that is, the financial loss is much larger than the operating cost or the total cost function is dominated by the financial loss, then, $n^* = \infty$;

3 Otherwise, $n^* > \mu$, and the value of n^* is the smallest integer satisfying the relationship $\mathrm{ETC}(n) < \mathrm{ETC}(n$ - $1)$, which yields

$$\begin{aligned} &\frac{n\mu}{(n-\mu)\phi_{[n-1]}} \times \left\{\left[\frac{(1+r)^n - 1}{r}\right]^{\alpha} - \left[\frac{(1+r)^{n-1} - 1}{r}\right]^{\alpha}\right\} \\ &\qquad < \frac{Le^{-\mu}\mu}{ky_s^{\alpha}} = \frac{L(1)}{C(1)} \end{aligned} \tag{5.11}$$

Clearly, the optimal number of the suppliers to be employed cannot be given in a closed form. Moreover, it is interesting to see that the right-hand-side of (5.11) is simply the ratio of the financial loss to the operating cost when only one supplier is used.

4.2 To Use or Not to Use Standby Suppliers?

The other strategy to reduce the supplier risks suggested in the literature is pooling resources, meaning that a network of suppliers is established so that if disaster strikes one, others will act as backups. We interpret one aspect of this decision-making situation as follows. During any given order cycle, a focal company orders items from its regular suppliers (RS's). However, on occasion, the RS's are not able to fulfill the complete order. The reasons for this are many, varying from a large-scale terrorist attack, or "act of God" (i.e., the weather, or a natural disaster), to an event unique to one or more of the RS's. One possible solution to cope with the issue of RS's not being able to fulfill a focal company's order, as suggested by the "network" concept, is for the focal company to engage a "Standby Supplier" (SS). If needed, the SS will supply the focal company with up to some requested number of items. Of course, the focal company must pay the SS some amount of money to "stand by," but, on the other hand, will be able to get its order fulfilled if a shortfall in the order from the RS's does occur. The "incremental cost of supply" will be either the money paid to the SS when he is not utilized, or the extra amount needed to be paid for the items needed when purchased on the spot market, rather than having been earlier contracted for.

Hence, a decision problem faced by the focal company naturally follows: (1) whether to employ a standby supplier; and (2) how many standby units should be established by contract. We study this decision using decision analysis and optimization approaches. Before proceeding, we present our notation and assumptions for modeling the decision problem.

Notation

Q = the order quantity – what the focal company needs and contracted for by the focal company with the RS's;

N = the number contracted for with the SS, a *decision variable;*

P = the unit price provided for in the contract between the focal company and RS's and SS;

I = the unit amount paid to the SS by the focal company to get an option on N items;

$P + E$ = the unit price for spot purchases that have not been contracted for with anyone;

X = the number of items unable to be supplied by the RS's, a *random variable.*

Assumptions

We assume that part of the focal company's contract with the SS is that if, in turn, the SS does not provide the requested amount, the SS must pay the focal company a sufficiently large penalty to virtually guarantee that the SS will, indeed, be prepared to fulfill any request the focal company makes (even if it means that the SS must, himself, have an 'SS'). We also assume that whenever the focal company makes a request to the SS, the price per unit is P, and for these units ordered, I is part of the price, P. Or, equivalently, the focal company need not pay the amount of I for any item actually ordered.

It is more straightforward to formulate the problem in terms of the cost incremental to $P \cdot Q$. Let $C(N, X)$ = the *incremental* cost, above and beyond $P \cdot Q$. Then,

$$C(N,X) = \begin{cases} E \cdot (X - N), & \text{if } X \geq N \\ I \cdot (N - X), & \text{if } N \geq X \end{cases}$$

The general case is

$$C(N,X) = \begin{cases} K_u \cdot (X - N), & \text{if } X \geq N \\ K_o \cdot (N - X), & \text{if } N \geq X \end{cases}$$

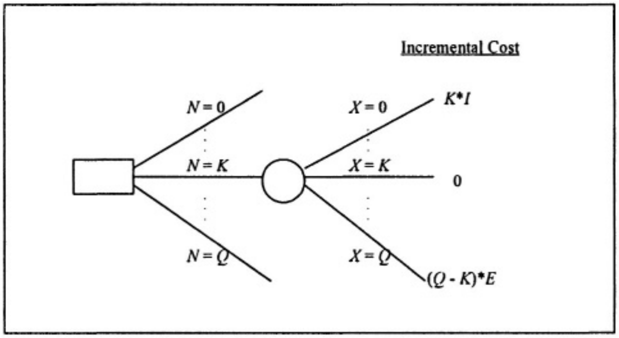
where K_u is the unit cost of "guessing under," and K_o the unit cost of "guessing over." This structure occurs frequently in decision problems, the classic perhaps being the so-called "Newsboy problem." It has been used extensively in inventory management (Silver, Pyke and Peterson, 1998). It is easy to see that if K_o is zero and $K_u \gg K_o$, X is chosen to be very large. The converse is a case where K_u is zero, or nearly zero, then $X = 0$.

The problem can be depicted in a decision-tree format as shown in Figure 5.4. If we let the density and distribution functions of the units of shortfall, X, be notated as follows:

$$f(K) = \mathrm{P}\{X = K\}$$
$$F(K) = \mathrm{P}\{X \leq K\}$$

then, it can be shown that the optimal value of N, the value that minimizes the expected value over X of $C(N, X)$, "EC(N)," is N^*, where,

$$F(N^*) = E/(E + I). \tag{5.12}$$



Notation:

N = # of standby units, 0, 1, ..., K, ..., Q; a decision variable
Q = order quantity placed by the focal company; a known parameter
X = # of units short from the order quantity, follows a probability distribution
I = unit amount paid to the Standby Supplier by the focal company to get an option on N standby units
E = extra cost per unit if the units short are purchased from spot market

Figure 5.4. A Decision-Tree Model for Using Standby Suppliers.

Should no N^* exist for which (5.12) holds exactly, which is usually the case when the probability distribution of X is discrete, and N^* must be an integer, then the optimal value of N is either of the following: (a) the value of N which makes the left hand side of (5.12) just exceed the right hand side of (5.12), or (b) one less than this integer value of N from (a).

The Shortfall Follows a Poisson Distribution

An initial question that may concern the focal company is under what conditions having a standby supplier is more cost effective than not having a standby supplier. We consider a scenario where the shortfall follows a Poisson distribution and then examine the property of the cost function. In general, the expected incremental cost function, EC(N), can be found after some algebra as follows:

$$\begin{aligned} \mathrm{EC}(N) &= I\sum_{i=0} N\left(\frac{\lambda_x^i e^{-\lambda_x}}{i!}\right)(N-i) + E\sum_{i=N+1}^{\infty}\left[(i-N)\left(\frac{\lambda_x^i e^{-\lambda_x}}{i!}\right)\right] \\ &= (I+E)NF_P(N) - EN + E\sum_{i=N+1}^{\infty} if_P(i) - i\sum_{i=0}^{N} if_P(i) \end{aligned}$$

$$= (I+E)NF_P(N) - E(N-\lambda_x) - (I+E)\sum_{i=0}^{N} if_P(i), \quad (5.13)$$

where $F_P(.)$ and $f_P(.)$ are the distribution and density functions of the Poisson variable (X), respectively. It is easy to see that when the number of standby units is set equal to zero, that is, when $N = 0$, then

$$\mathrm{EC}(0) = E \cdot \lambda_x. \quad (5.14)$$

Furthermore, if one standby unit is used, that is, $N = 1$, it is straightforward to see that

$$\begin{aligned} \mathrm{EC}(1) &= (I+E)P\{X \leq 1\} - E(1-\lambda_x) - (I+E)P\{X=1\} \\ &= (I+E)P\{X=0\} - E(1-\lambda_x) \\ &= (I+E) \cdot e^{-\lambda_x} - E(1-\lambda_x) \quad (5.15) \end{aligned}$$

Hence, the condition under which employing a standby supplier (i.e., $N^* \geq 1$) is better than no standby supplier (i.e., $N^* = 0$) is

$$EC(1) - EC(0) < 0. \quad (5.16)$$

Substituting (5.14) and (5.15) into (5.16), we obtain the condition as follows:

$$(I+E) \cdot e^{-\lambda_x} - E < 0,$$

which further implies that

$$\lambda_x > -\ln\left[E/(I+E)\right]. \quad (5.17)$$

Therefore, we conclude that whether the focal company should employ a standby supplier depends on the relationship in (5.17) between the mean shortfall (λ_x) and the ratio of the extra cost per unit if the shortfall is obtained from the spot market (E) to the sum of this cost and the unit amount paid to the SS by the focal company to get an option of having standby units (I).

It is seen that the decision-tree approach proves again an effective modeling technique to examine the decision on the optimal use of the standby suppliers. The key parameters of the model and the final solution include the probability distribution of the shortfall experienced by the buying firm, the cost of keeping a standby supplier, and the extra unit cost if the unsatisfied items are obtained from the spot market.

4.3 Observations and Future Research Directions

We have relied on one of the famous decision-making under uncertainty techniques, namely decision trees, to model and examine two important methods for dealing with the supplier risks suggested in the literature: one is the use of several suppliers and the other is the establishment of standby suppliers. In the first case, the cost factors considered include the financial loss caused by disasters of all suppliers down and the operating cost of the buying firm doing business with multiple suppliers. In the other case, we have examined the merit of using standby suppliers if the regular suppliers cannot satisfy all demands by looking at the probability of the unmet demand, the extra cost incurred if the unsatisfied items are obtained from spot market, and the cost of employing a standby supplier. In both cases, the probabilities of the occurrence of certain supplier risks and cost trade-offs are captured nicely by decision trees, enabling us to find the expected cost functions and then the optimal decisions.

We also need to point out a number of limitations of the decision-tree based optimization models we have proposed above. When studying the decision of the best number of suppliers to be used in the basic scenario, we have assumed that the probability of the unique event that brings down a particular supplier is the same for all suppliers. This assumption has greatly simplified the analysis; however, they need to be relaxed if the supply base is truly diverse, in other words, the suppliers are heterogeneous, possessing different probabilities of being down. Although in the other scenario we used more realistic functions to examine the decisions based on the operating cost and financial loss, more techniques are needed for modeling the unavailable number of suppliers. Furthermore, the decision-making problem is studied in a single period. For future studies, the two cost elements should be modeled as time-dependent discounted factors.

In the other decision of whether to use standby suppliers is concerned, one of our key assumptions is that the order quantity of the focal company is a known constant, and that only the unmet demand to be satisfied by the standby supplier is a random variable. We think that this assumption can be relaxed to more realistically capture the real decision-making situation, that is, the demand to be contracted with the regular suppliers is not known beforehand, and thus also a random variable. Additionally, a multi-period analysis of the decision-making problem calls for attention.

In addition to the above areas for future studies, we believe that the issues associated with supply risk management open up numerous opportunities for management science or operations research professionals. The commonly used methods for dealing with the risks of suppliers or supply market are summarized and strategically discussed in the literature, but quantitative models that can aid and guide risk managers decision-making are sparse. We think that questions as to "what strategies" are available to reduce the supplier risks are well answered in the literature, but it is "how" and "when" to use these strategies that still require a great deal of research interests and efforts, especially for strategies such as how to select the low-risk suppliers, how to diversify the supply base, how and when to stockpile, and how many resources to pool. Quantitative models that can be used to examine the efficiencies and the impact of these strategies are in great need.

5. Concluding Remarks

Today's supply chains are becoming not only more efficient with the aid of advanced information technologies, but also riskier due to the heavy inter-dependence of numerous chain links. These links are subject to interruptions, disruptions or disasters, and a breakdown at any point in the continuum may defeat the ultimate goal of the supply chain partners. Although many studies focusing on business risks in various contexts have been presented in the literature over the years, research effort devoted to understanding the risks associated with suppliers and the supply market has been limited, especially from a quantitative aspect. As more companies' supply chains are becoming more supplier-dependent and the physical supply cost can easily account for thirty percent of the total cost of goods sold, it is even more important and critical for corporations to understand and manage potential supply-side risks.

In this chapter, we first, through extensive literature review, present a taxonomic review of supply-side risk types, a four-step supply risk management process, and a list of techniques that help accomplish each step. These managerial aspects of supply risk management should provide guidelines and starting points for supply chain managers. Then, from a quantitative point of view, we study two strategies suggested in the literature for dealing with certain supplier risks, namely the use of multiple sources and the development of a supply network. In particular, we propose two optimization-based decision tree models that can effectively formulate the decision-making situations in which the questions of how many suppliers should be used and whether to use standby suppliers are addressed. Finally, we have suggested areas for possibly expanding our current models and future research directions for studying some well-known supply risk management strategies.

References

Ballou, R.H. (1999), Business Logistics Management: Planning, Organizing, and Controlling the Supply Chain, 4/e, Prentice Hall, Upper Saddle River, New Jersey 07458.

Berger, P.D., Gerstenfeld, A., and Zeng, A.Z. (2002a), "How Many Suppliers Are Best? A Decision-Analysis Approach," working paper series #02042002, Department of Management, Worcester Polytechnic Institute.

Berger, P.D., Gerstenfeld, A., and Zeng, A.Z. (2002b), "The Optimal Use of Standby Suppliers: A Decision-Analysis Approach," working paper series #06042002, Department of Management, Worcester Polytechnic Institute.

Clausen J., Hansen, J., Larson, J., and Larson, A. (2001), "Disruption Management," OR/MS Today, October Issue, pp. 40-43.

Cox, J.F. III, and Blackstone, J.H. Jr. (1998), APICS Dictionary, 9/e, APICS–the Educational Society for Resource Management, Stock No. 01102.

Dornier, P-P., Ernst, R., Fender, M. and Kouvelis, P. (1998), Global Operations and Logistics: Text and Cases, John Wiley & Sons, Inc.

Eppen, G.D., Gould, F.J., Schmidt, C.P., Moor, J.H., and Weatherford, L.R. (1998), Introduction to Management Science, 5/e, Prentice Hall, Inc.

Gilbert, G.A. and Gips, M.A. (2000), "Supply-Side Contingency Planning," Security Management, March Issue, pp. 70-74.

Grey, W. and Shi, D. (2001), "Value Chain Risk Management," presentation at the INFORMS 2001 Miami Meeting, Session MA23, November 4-7, Miami Beach, FL.

Impetus Training: Scenario Analysis, `www.impetustraining.co.uk/scenario.htm#Concept`

Johnson, M.E. (2001), "Learning from Toys: Lessons in Managing Supply Chain Risk from the Toy Industry," California Management Review, 43(3), pp. 106-124.

Michalski, L. (2000), "How to Identify Vendor Risk," Pharmaceutical Technology, October Issue, pp. 180-184.

Nahmias, S. (2001), Production and Operations Analysis, 4^{th} edition, McGraw-Hill Companies, Inc.

Process Mapping and Alignment: http://www.cranfield.ac.uk/coa/macro/nextgen/newpage31.htm

Silver, E.A., Pyke, D.F., and Peterson, R. (1998), Inventory Management and Production Planning and Scheduling, 3^{rd} edition, John Wiley & Sons, Inc.

Smeltzer, L.R. and Siferd, S.P. (1998), "Proactive Supply Management: The Management of Risk," International Journal of Purchasing and Materials Management, 34(1), Winter Issue, pp. 38-45.

Souter, G. (2000), "Risks from Supply Chain Also Demand Attention," Business Insurance, 34(20), pp. 26-28.

Zeng, A.Z. and Berger, P.D. (2003), "Single versus Multiple Sourcing in the Presence of Risk and Uncertainty," working paper series #07312003, Department of Management, Worcester Polytechnic Institute.

Zsidisin, G.A. and Ellram, L.M. (1999), "Supply Risk Assessment Analysis," PRACTIX: Best Practices in Purchasing & Supply Chain Management, 2(2), pp. 9-12.

Zsidisin, G.A., Panelli, A. and Upton, R. (2000), "Purchasing Organization Involvement in Risk Assessments, Contingency Plans, and Risk Management: An Exploratory Study," Supply Chain Management: An International Journal, 5(2), pp. 187-197.

Zsidisin, G. (2001), "Measuring Supply Risk: An Example from Europe," PRACTIX: Best Practices in Purchasing & Supply Chain Management, 4(3), June 2001, pp. 1-6.

Chapter 6

DEMAND PROPAGATION IN ERP INTEGRATED ASSEMBLY SUPPLY CHAINS: THEORETICAL MODELS AND EMPIRICAL RESULTS

S. David Wu
Department of Industrial and Systems Engineering
Lehigh University
Bethlehem, Pennsylvania
david.wu@lehigh.edu

Mary J. Meixell
School of Management
George Mason University
Fairfax, Virginia
mmeixell@gmu.edu

Abstract This chapter studies supply chain demand propagation in an ERP-integrated manufacturing environment, where item demands are interrelated by their assembly structure. The integrated planning environment is commonplace in industries such as automotive and electronics, where detailed production and material requirements decisions are communicated electronically between facilities and then refreshed frequently. While tightly integrated supply chains have clear benefits, distortion can occur in the demand signals between the facilities that assemble end-items and those that manufacture sub-assemblies and components. Using both analytical and empirical tools, we explore an ERP model using basic lot-sizing logic under fairly general settings. We examine key factors that influence demand variation in the assembly supply chain, assess their effects, and develop insight into the underlying supply processes. We find that (1) order batching by downstream facilities plays a principal role in upstream demand amplification, (2) the commonly used schedule release policy in ERP systems may cause unnecessary nervousness in the supply chain, (3) the interplay of component sharing and demand correlation may contribute significantly to upstream demand fluctuation, and (4) tightly set production capacity may help to dampen demand variations, while paying a price on increased supply chain inventory. We also find that end-item demand variation, a commonly believed source of uncertainty, has no significant impact on overall demand amplification under the conditions studied here.

1. Introduction

Demand propagation – the translation of orders through the tiers in a supply chain – is a fundamental characteristic and a primary performance driver in supply chains. The demand pattern for a particular end-product is altered as orders are processed into production schedules at successive facilities, often distorting the original pattern. This distortion can adversely affect cost and delivery performance throughout the supply chain. One well-known type of distortion is the bullwhip effect, where the variation in order quantities amplifies from tier to tier. Amplification in demand makes production scheduling, inventory control, and transportation planning increasingly difficult to manage and can cause suppliers to miss due dates. To avoid these adverse effects of demand propagation, many manufacturers have sought to eliminate distortion by improving the information availability along their supply chains. First through EDI, and more recently with web-integrated ERP portals, information has been made both more accurate and timely.

This research is motivated by our experience in analyzing a supply chain in a major automotive company during the late 1990's. In this supply chain, demand is communicated along the chain as electronically posted production schedules and the related orders for material as determined by that product's bill of materials. The supplier views a "release" that represents the customer's orders at that particular point in time, uses this information to generate their own production schedules, and then post their schedules and related material requirements for the next supply tier to use. As sophisticated as they are, these systems are not without problems. ERP-driven material schedules change from one release to the next, frequently and sometimes erratically. In automotive supply chains, daily production and material schedules are maintained in systems that typically use a three-week planning horizon, with updates to the schedule several times a day. While the frequent update attempts to maintain concurrency and synchronization throughout the chain, it often increases demand amplification and creates unintended *nervousness* in the system. This nervousness may be attributed to the practice of updating the schedule based on the most recent schedule changes of the immediate customer and to the accumulation in time lag along the chain. System nervousness translates into significant operating costs for the supply chain, in the form of safety stock inventory, expedited transportation cost, or unplanned production overtime.

Investigation into the fundamental behavior of demand has enhanced managerial aptitude for avoiding undesirable supply chain behavior. The work by Lee et al., 1997, identifies the causes and suggests mitigating actions to eliminate the distortion for multi-period inventory systems operated under a periodic review policy. The approach of studying basic supply chain phenomena under a generalized setting (cf. Sterman, 1989; Slats et al., 1995; Lee and Billington, 1993) is beneficial because it provides managerial insights under mild assumptions. Another line of research concentrates on identifying design or operational policies that help to avoid undesired supply chain behavior (Takahashi et al., 1987; Towill, 1992; Lee, 1996).

The issues associated with integrating decisions along a supply chain have been addressed for applications in retail, transportation logistics, distribution channel design, and service part logistics. This chapter, however, addresses issues that exist in the production pipelines of supply chains integrated by ERP systems. Ganeshan et al., 1999, proposes a research taxonomy that categorizes this research as focusing on operational efficiency in the area of production planning and scheduling. Other research in this same category addresses different aspects of the problem. Graves et al., 1998, models requirement planning using a multi-echelon inventory system. Ertogral and Wu, 2000, proposes an auction-theoretic approach to the coordination of production planning in the supply chain. Kruger, 1997, Karabuk and Wu, 2002, and Karabuk and Wu, 2003, focus on the integration of marketing, operations, and procurement decisions in supply chain planning. Lederer and Li, 1997, addresses the issue of pricing and delivery time competition among firms in the supply chain. They define and compute competitive equilibrium in this environment under different cases. Levy, 1997, focuses on geographically dispersed global supply chains, with some discussion on planning and scheduling issues. O'Brien and Head, 1995, considers the environment of EDI-integrated supply chains, but their focus is on the establishment of efficient business processes that would facilitate the JIT operational environment. Kekre et al., 1999, reports field studies on the role of EDI in JIT supply chains, but their study focuses on the processing and payment of orders, rather than production coordination.

The ERP model we use to express the basic lot-sizing logic derives from the multi-level lot-sizing literature, which supplies a decision model of the underlying material requirements planning process. Reviews of this literature are available (cf. Bahl et al., 1987; Maes and Wassenhove, 1988; Nam and Logendran, 1992; Kimms, 1997). Although most of these models are intended for facility-level production planning, they have been generalized more recently to handle broader scope of production planning (Wu and Golbasi, 2003; Balakrishnan and Geunes, 2000; Vidal and Goetschalckx, 1997; Thomas and Griffin, 1996; Bhatnagar et al., 1993). Several researchers focus on the value of information and information technology in supply chain operations (Chen et al., 1999; Camm et al., 1997; Geoffrion and Power, 1995; and Singh, 1996). Srinivasan et al., 1994, examines the behavior and performance impact on vertically information integrated supply chains. There are also documented industry cases (Lee and Billington, 1995; Martin et al., 1993; Robinson, Jr. et al., 1993; Arntzen et al., 1995), which examine the role of information in supply chain implementations.

2. The Assembly Supply Chain Environment

We view an assembly supply chain as a network of facilities $k = 1, \ldots, K$ and processes structured to manufacture a variety of products, $i = 1, \ldots, N$. The products consist of end-items and their components, to be produced in multiple facilities over multiple time periods. Each end-item has a bill of material (BOM) described by a product structure. The supply chain has a supply structure where a set of facilities is setup to produce each item described in the product structure. Multiple supply tiers are defined by the product structure. Figure 6.1 illustrates the assembly structure and the supply structure. The *assembly structure* can be represented by an $n \times n$ matrix $[a_{ji}]$, where a_{ji} represents the number of units of component j required for the production of one unit of item i. The final products for the supply chain are presented as the first-tier items in the chain, and customer sales for that item make up the demand. There can be multiple tiers of subassemblies or components in the network, organized by levels each is removed from the final product at the top tier. The first tier products are shipped directly to the market, the second tier ships to first tier, and so on. Each product in each tier has its own set of components, following the form of its BOM structure. A component j may be shared by multiple products i in several different BOM structures (e.g., 7, 10, 11 in facility III). Thus, the planned production x_{jt} for an upper tier product i imposes a demand $(a_{ji}x_{it})$ on its lower tier components j.

A delivery is made strictly based on the customer's demand as described in the schedule, i.e., a planned production x_{it} for product i triggers a series of orders of size $a_{ji}x_{it}$ to the designated supplier of each

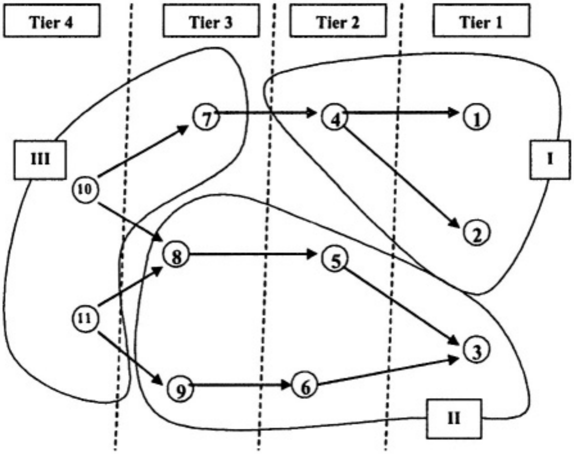


Figure 6.1. An assembly supply chain with 3 end-items (1,2 and 3), 8 components and sub-assemblies (4 to 11), 3 production facilities (I to III), and 4 supply tiers.

component j. Under the above operational assumptions, each schedule along the supply chain depends on the schedule at the previous tier.

2.1 Defining demand amplification in the supply chain

Numerous effects influence the propagation of demand through a supply chain. Consider an item i and its component j. The production lot-size established for item i imposes a consolidated demand quantity for its component j. The policy under which these lot-sizes are determined at each facility determines, to a large extent, the *demand* behavior in a supply chain. We define three measures of demand amplification that are analytically tractable and sufficient to characterize demand variations across supply tiers. The first is the degree to which demand varies from an item to its immediate component within a single schedule release (e.g., between items 1 and 4, or 4 and 7, or 7 and 10, etc. in Figure 6.1).

DEFINITION 6.1 **(Single Release Item-Component Demand Amplification, DA^s_{ij})** *For an item i and its immediate component j, item-component demand amplification, DA^s_{ij}, is the change in the coefficient of variation between item i demand and component j demand in a single schedule release. Specifically,*

$$DA^s_{ij} = CV(X) - CV(D)$$

where random variable D represents the demand of item i over periods $t = L_i + 1, \ldots, T$, and random variable X represents non-zero internal demand generated from D over periods $t = 1, \ldots, T - L_i$, using some lot-sizing policy.

Thus, DA^s is the most basic and direct measure of demand variation between an item and its component within one single schedule release. Since a schedule release covers multiple time periods, DA^s pertains to the batching of orders across the planning horizon. In an ERP-integrated assembly environment, production schedules are updated frequently via multiple schedule releases. We define a second type of demand amplification that applies to this environment.

DEFINITION 6.2 **(Multiple Release Item-Component Demand Amplification, DA_{ij}^m)** *For an item i and its immediate component j, item-component demand amplification, DA_{ij}^m, is the change in coefficient of variation between item i demand and component j demand across multiple schedule releases. Specifically,*

$$DA_{ij}^m = CV(X^m) - CV(D^m)$$

where random variable D^m represents the demand of item i over periods $t = L_i + 1, \ldots, T$, in schedule releases $\tau = 1, \ldots, \Theta$, and random variable X^m represents non-zero internal demand generated from D over periods $t = 1, \ldots, T - L_i$, in corresponding releases $\tau = 1, \ldots, \Theta$, using some lot-sizing policy.

Thus, DA^m measures the demand amplification between an item and a single component over multiple schedule releases. Similar to DA^s, DA^m pertains to multiple periods in the planning horizon. In this nomenclature, D^m is a vector-valued random variable that represents the observed demand for an item i over multiple releases τ. Likewise, X^m is a vector-valued random variable for the computed requirements for the component j also over multiple releases τ.

While these first two measures quantify change across an item-component pair, a third measure quantifies the change in demand variation over multiple item-component pairs across two adjacent supply tiers. Consider a segment of a supply chain that is comprised of two supply-tiers k and l where a set of items are produced in tier k with components produced in tier l (e.g., items 4, 5, 6 in tier 2 and 7, 8, 9 in tier 3). We define a third type of demand amplification as follows.

DEFINITION 6.3 **(Tier-to-Tier Demand Amplification, TA_{kl})** *For supply tiers k and l, tier-to-tier demand amplification TA_{kl}, is the change* *in coefficient of variation between tier k and tier l in a single schedule release. Specifically,*

$$TA_{kl} = CV(X^l) - CV(D^k)$$

where random variable D^k represents the demand of all items at tier k over all periods $t = 1, \dots, T$, and random variable X^l represents the demand of all items at tier l over all periods $t = 1, \dots, T$.

3. Analyzing Demand Propagation

3.1 Order batching

Order batching is a means to consolidate production demand for the purpose of reducing setups. In supply chains with multiple manufacturing tiers, order batching in an upper tier facility may amplify demand variation at a lower tier. We analyze this simple but commonly seen effect in this section. We will separate the effect of order batching from that of capacity, which is examined in Section 3.4. For any facility in the supply chain that has linear setup (K_i) and inventory-holding costs (h_i), we may characterize its scheduling system by an objective that trades off set-up and inventory holding costs as follows:

$$z = \min \sum_{i=1}^{N} \sum_{t=1}^{T} (h_i y_{it} + C_{it} x_{it}) \text{ where } C_{it} = \begin{cases} c_{it} x_{it} + K_i & \text{if } x_{it} > 0 \\ 0 & \text{if } x_{it} = 0. \end{cases}$$

This results in a nonnegative order size of $(a_{ji} x_{jt} + \delta_{jt}^{+} - \delta_{jt}^{-})^{+}$ where δ_{jt}^{+} represents a positive adjustment (padding) for item j in period t and δ_{jt}^{-} represents a negative adjustment (shrinking). The lot scheduling problem is subject to the inventory balance requirement as follows,

$$y_{j,t-1} + x_{j,t-L_j} - y_{jt} = \sum_{i=1}^{N} (a_{ji} x_{it}^{\tau} + \delta_{it}^{+} - \delta_{it}^{-}) + r_{jt} \qquad \forall j, t$$

where r represents external demands for the component. If all constraints are linear, such scheduling systems have the following well-known properties.

LEMMA 6.4 *The inventory balance requirement is a Leontief substitution system (Veinott, Jr., 1969) as long as the gross adjustment made through the supply tiers create nonnegative internal demand, i.e., $r_{it} + \sum_{j=1}^{N} (\delta_{jt}^{+} - \delta_{jt}^{-}) \geq 0$ and $\sum_{j=1}^{N} (a_{ij} x_{jt} + \delta_{jt}^{+} - \delta_{jt}^{-}) \geq 0$.*

3.1.1 Demand amplification between an item and a component. A basic building block of demand variation in a supply chain is the way the demand of an item is passed to its components. If the orders of an item are not batched to reduce setup, then the item's demand can simply be passed on to its components, after floating for lead time, regardless of the assembly structure a_{ji}.

PROPOSITION 6.5 *Suppose the production scheduling system in the supply chain satisfies the Leontief property. Thus, when the inventory costs dominate the setup costs, there will be no single-release item-component demand amplification, i.e.,* $DA^s_{ij} = 0$.

The *Leontief* property (i.e., $x_{t-L_j} \cdot y_{t-1} = 0$) states that when capacity is not binding and the inventory cost dominates the setup cost, a facility will build each order as needed and carry no extra inventory ($x_{t-L_i} > 0$ and $y_{t-1} = 0$). As a result, the demand for the end item, r_{it}, translates through the supply chain, based on the assembly structure matrix $[a_{ji}]$, un-altered by production scheduling, and becomes the internal demand for a component, $x_{j,t-L_i}$. In the more general case when the inventory costs do not dominate the setup costs at all times, the Leontief property implies that some facility will consolidate orders and build ahead, (i.e., $y_{t-1} > 0$ and $x_{t-L_i} = 0$). Suppose v is the number of periods where orders are consolidated for production. The following proposition addresses where the demands for item i are independent.

PROPOSITION 6.6 *Suppose* v *periods of orders for component* j *are batched to satisfy internal demand imposed by item* i, *i.e.,* $x_{j,t-L_i} = a_{ji}(r_{it} + r_{i,t+1} + \cdots + r_{i,t+v-1})$, *for periods* $t = L_i + 1, \ldots, T - L_i$. *If the demand stream* $r_{i1}, \ldots, r_{iT}$ *are drawn from i.i.d. random variables, then*

(1) component j*'s order size increases by a factor of* v*, with mean* $a_{ji}v$ *and variance* $a^2_{ji}v$

(2) the item-component demand amplification DA^s_{ij} *decreases by a factor of* $\left(\frac{1}{\sqrt{v}} - 1\right)$.

See the Appendix for the proofs of propositions and corollaries in this chapter.

If the demands are independent, increasing the batch size v in any production period increases the mean and variance of the order size at the component level. However, since both the mean and variance increase by a factor of v, the item-component demand amplification actually decreases as the lot size increases. This is not true, however, when the demands are correlated.

COROLLARY 6.7 *Suppose the demands $r_{i1}, \ldots, r_{iT}$ are drawn from correlated random variables with a coefficient of correlation ρ and the same variance, then the following are true after batching:*

(1) item-component demand amplification DA^s_{ij} changes by a factor of $\left(\sqrt{\frac{1+v\rho-\rho}{v}}-1\right)$

(2) when $\rho = 1$, there is no item-component demand amplification, i.e., $DA^s_{ij} = 0$.

It should be noted that when $v > 2$ and ρ is negative such that $v\rho - \rho > -1$, the factor to DA^s becomes problematic because of the negative sign under the radical. For example, when $\rho = -\frac{1}{2}$ and $v = 4$, the factor cannot be computed. Also, it is important to note that DA^s_{ij} is defined such that we compute the coefficient of variation only over the periods where component j has *non-zero production.* All periods with no production of j after batching are dropped from the calculation. The random variable *X* is a simple summation of random variable *D* every v periods, and we are only computing the variance and the *CV* of *X*. The more general case, which considers all items over all periods, is given in the following.

3.1.2 Tier-to-tier demand amplification over multiple items.

We now consider the multiple-item cases where each production period may be populated with non-zero production of different components. We would like to know the *collective demand amplification* of all items in a particular supply *tier.* To do this, we first need to know for each item the demand variation of its production over *all* zero and non-zero production periods. Suppose the batch size of all items is v, consider a particular item-component pair (i, j). Denote $x_{j1}, x_{j2}, \ldots, x_{jT}$ the demand stream for component j drawn from random variables $Y_1, \ldots, Y_T$, and $r_{i1}, r_{i2}, \ldots, r_{jT}$ the demand stream for item i drawn from random variables $D_1, \ldots, D_T$, respectively. For any block of v periods in the planning horizon 1,... ,*T,* there is exactly one *non-zero production period* for component j, i.e., $Y_t = a_{ji}E(D_s + \cdots + D_{s+v})$, $t < s$, and all other periods in the block do not produce j, or $Y_t = 0$. We define random variable *Y* as follows:

$$Y = \begin{cases} a_{ji}E(D_1 + \cdots + D_v) & \text{with probability } \frac{1}{v} \\ 0 & \text{with probability } 1 - \frac{1}{v}. \end{cases}$$

Suppose each random variable $D_1, \ldots, D_T$ has an expected value *E*(*D*). Then, the random variable *Y* has expected value $E(Y) = a_{ji}E(D)$. Note that the component demand Y_t is defined based on the *expected* batch demand when there is non-zero production, not the actual realization of random variables $D_s, \ldots, D_{s+v}$. This is because the components are built in advance $(t < s)$, before the actual item demands are known. The only demand information available at that point in time is the expected value. We now present the following lemma:

LEMMA 6.8 *Under the above setting, suppose the batch size is v and each production period within a v-period block has equal probability $\frac{1}{v}$ to be chosen for producing one batch of j. Then:*

(1) the variance for random variable Y is $(v-1)E(Y)^2$

(2) the coefficient of variation $CV(Y) = \sqrt{v-1}$.

From Lemma 6.8 we can conclude the following for a single item.

PROPOSITION 6.9 *If the demands $r_{i1}, r_{i2}, \ldots, r_{jT}$ are drawn from i.i.d. random variables, then the tier-to-tier demand amplification is an increasing function of the lot size as follows:*

$$\frac{(v-1)\sqrt{v-1}}{v}.$$

Now consider N item-component pairs. Denote $D1, \ldots, DN$ random variables characterizing the item demands, and $Y1, \ldots, YN$ characterizing the corresponding component production over the planning periods. Thus the production orders in the next tier for all components is the sum of random variables $Y1 + \cdots + YN$. Suppose for each component $\mu = E(Yi) = E(Di)$. Thus $E(Y1 + \cdots + YN) = N\mu$. The following results suggest that the demand variation increases from one supply tier to another as a function of the lot size v.

PROPOSITION 6.10 *Under the above setting, if the item demands for all N items are i.i.d. random variables, then the following are true at the component tier:*

(1) the variance of production orders is $N(v-1)\mu^2$

(2) the coefficient of variation is $\sqrt{\frac{v-1}{N}}$

(3) the tier-to-tier demand amplification is an increasing function of the lot size while a decreasing function of N as follows:
$\sqrt{\frac{v-1}{N}}\left(1 - \frac{1}{\sqrt{v}}\right)$.

Fortunately, this amplification effect reduces when an increasing number of items are processed simultaneously, as is the case with larger *N*. Nonetheless, the number of items that could be realistically processed at the same time is limited by setup capability in the facility, and the need to group items with similar assembly processes.

3.2 Multiple schedule releases and supply chain nervousness

As discussed previously, the automotive supply chain tends to update production schedules frequently, driving an exaggerated level of demand variance relative to true demand variance. This nervousness can be attributed to two sources: (1) *myopic update policy,* a common policy for ERP systems that updates the schedule based on the *most recent* schedule changes of the immediate customer, and (2) *accumulated time-lag,* the time delay between any customer-supplier scheduling updates, which propagates and accumulates along the chain. The latter contributes to system nervousness when a considerable gap develops between the upstream demand view and the end-item demand. The errors must be corrected periodically by various "re-synchronizing" attempts, which lead to erratic schedule changes in upstream facilities, e.g., an item originally scheduled for a high-volume production may be suddenly cancelled while an unplanned item takes its place. In the following, we present an analysis on this issue to show that the current practice on schedule releases creates an unnecessarily more nervous scheduling system since it introduces higher demand variance between each customer-supplier pair.

Consider multiple schedule releases $\tau = 1, \dots, \Theta$ over multiple production periods $t = 1, \dots, T$. For each schedule release τ, let r_{it}^{τ} be the demand of item i in period t, and $x_{j,t-L_i}^{\tau}$ be the internal demand imposed on i's component j in period $t - L_i$. After order batching, v periods of orders for component j are combined to satisfy the demand of item i, i.e., $x_{j,t-L_i}^{\tau} = a_{ji}(r_{it}^{\tau} + r_{i,t+1}^{\tau} + \cdots + r_{i,t+v}^{\tau})$, the order size and demand amplification. We first present a straightforward result, which states that when the multiple schedule releases are mutually independent, item-component demand amplification behaves the same way as in the single release cases.

PROPOSITION 6.11 *Suppose as a result of batching that v periods of orders for component j are consolidated to satisfy the internal demand imposed by item i, i.e., $x_{j,t-L_i}^{\tau} = a_{ji}(r_{it}^{\tau} + r_{i,t+1}^{\tau} + \cdots + r_{i,t+v}^{\tau})$, for $t = L_i + 1, \dots, T - L_i$ and $\tau = 1, \dots, \Theta$. If the demands $r_{i1}^{1}, \dots, r_{iT}^{1}, \dots, r_{i1}^{\Theta}, \dots,$ r_{iT}^{Θ} are drawn from i.i.d. random variables, then*

(1) j's order size increases by a factor of v, with mean $a_{ji}v$ and variance $a_{ji}^2 v$, but

(2) the item-component demand amplification DA_{ij}^m decreases by a factor of $\left(\frac{1}{\sqrt{v}} - 1\right)$.

The above result is intuitive except that schedule releases are rarely independent in practice. If we are to consider the nature of dependency among multiple schedule releases, we must consider the policies that the ERP system uses to update these releases. Procurement practices typically require that suppliers provide component materials as specified by the *most recent* schedule release. This can be described more precisely as a "freeze-up-to" policy – at the end of period $(t-k)$, freeze the period-t customer schedule as the "most recent" schedule that defines the internal demand for period t, where k represents the time lag between the two systems. Similarly, the release in period $t-k+1$ is used to determine period $t+1$ demands, and $T-k$ is used for period T, etc. Since the schedule is updated at least once in each planning period, only one period from each released schedule is actually implemented while the remainders are discarded. Consider each order release for item i as a vector-valued random variable as follows:

$$D^\tau = \begin{bmatrix} D_1^\tau \\ \vdots \\ D_T^\tau \end{bmatrix} \qquad \text{for } \tau = 1, \ldots, \Theta.$$

After batching v periods of orders for production, this order release, the customer's production and material requirement schedule, is translated into scheduling releases as follows:

$$X^\tau = \begin{bmatrix} D_1^\tau + \cdots + D_v^\tau \\ \vdots \\ D_{T-v+1}^\tau + \cdots + D_T^\tau \end{bmatrix} = \begin{bmatrix} X_1^\tau \\ \vdots \\ X_{T-v+1}^\tau \end{bmatrix} \qquad \text{for } \tau = 1, \ldots, \Theta.$$

Each vector-valued random variable X^τ represent a schedule release. If we write all the X^τ 's together over multiple releases, we have a matrix as follows:

$$M_{(T-v+1)\times\Theta} = \begin{bmatrix} X_1^1 & & \cdots & & X_1^\Theta \\ & X_{v+1}^2 & & & \\ \vdots & & X_{2v+1}^3 & & \\ & & & \ddots & \\ X_{T-v+1}^1 & & \cdots & & X_{T-v+1}^\Theta \end{bmatrix}.$$

Suppose a release index τ is always k periods ahead of the time index, i.e., $t = \tau + k$. Then we may represent the schedule using the above "freeze-up-to" policy as the diagonal terms of matrix M which is a random vector as follows:

$$S = \begin{bmatrix} X_1^1 \\ X_{v+1}^2 \\ \vdots \\ X_{T-v+1}^{\Theta} \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_{T/v} \end{bmatrix}.$$

In a typical automotive environment where a production schedule is updated frequently, the schedule frozen at time $(t-k)$ is not truly the "most recent" schedule, because additional schedule changes may occur during the production lead time *L*. The discrepancy is absorbed by the suppliers' safety stock. For simplicity, we may assume the level of supplier's safety stock as a non-decreasing function of the customer's demand variance. Since in reality there will be always discrepancies between the "frozen" customer schedule and the true "most recent" schedule, we consider an alternative schedule update policy for the supplier. Instead of defining the customer demand based the *last* schedule released before time $(t - k)$, compute the customer demand as the *expected quantity of all releases up to* $(t - k)$. In other words, we compute from matrix *M* a vector

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_{v+1} \\ \vdots \\ \bar{X}_{T-v+1} \end{bmatrix}$$

where

$$\begin{aligned}
\bar{X}_1 &= X_1^1 \\
\bar{X}_{v+1} &= \frac{1}{v}(X_{v+1}^1 + X_{v+1}^2 + \cdots + X_{v+1}^v) \\
\bar{X}_{2v+1} &= \frac{1}{2v}(X_{2v+1}^1 + \cdots + X_{2v+1}^v + X_{2v+1}^{v+1} + \cdots + X_{2v+1}^{2v}) \\
&\vdots \\
\bar{X}_{T-v+1} &= \frac{1}{T-v}(X_{T-v+1}^1 + \cdots + X_{T-v+1}^v + \cdots + \\
&\qquad X_{T-v+1}^{v+1} + \cdots + X_{T-v+1}^{T-v}).
\end{aligned}$$

Using the above policy as a benchmark, we now present the following important result.

PROPOSITION 6.12 *Given multiple schedule releases from an upper tier customer represented by matrix M, and the freeze-up-to schedule represented by random vector S with a variance of* Var(*S*). *Then an alternative policy that uses the expected quantity of all up-to-date releases (as defined by vector $\bar{X}$) will reduce the variance by a factor of v for all but the first period.*

The above result suggests that the practice of updating the supplier's schedule following the "most recent" customer schedule creates demand variance that is v times higher than one that updates the schedule using the expected quantities of up-to-date releases. Hence, more supply chain nervousness and higher safety stock levels results if the most recent release is followed. Intuitively, if one views a given customer schedule release as merely an instance in a *population* of possible releases, where an underlying distribution is known, any method that attempts to estimate the underlying distribution reduces the variance.

3.3 Component commonality

Up to this point, we have focused our attention on the demand amplification between particular item-component pairs. In a simple assembly structure, where each component feeds exactly one item, the above analysis can be readily generalized to multiple tiers. However, in general assembly structures, where a component may feed into multiple items, the extension is not as straightforward. The demands of multiple upper tier items may have a joint effect on components in a lower tier. This particular issue is important since automotive supply chains often consolidate their product designs to achieve a high-level of component sharing. We examine this effect in the context of order batching as established earlier.

Using the same notation, define the simple assembly structure as one where the production of component j, say x_{jt}, will be used to satisfy the demand for exactly one type of item i. That is, $x_{j,t-L_i} = a_{ji}(r_{it}+r_{i,t+1}+\cdots+r_{i,t+v-1})$, for $t = L_i+1,\ldots,T-L_i$, where $x_{j1}, x_{j,v+1},\ldots,x_{j,T-v+1}$, is the realization of random variables $X_1,\ldots,X_{T-v+1}$, and $r_{i1}, r_{i2},\ldots,r_T$ is that of random variables $D_1,\ldots,D_T$. We know from Proposition 6.6 that if item demands $D_1,\ldots,D_T$ are i.i.d., $E(X) = a_{ji}vE(D)$ and $Var(X) = a_{ji}^2 v\, Var(D)$. On the other hand, in a general assembly structure, the production of component j, say x_{jt}, will be used to satisfy the demand for multiple items $i_1,\ldots,i_n$, for each item i. In other words,

$$x_{j,t-L} = a_{ji_1}(r_{i_1t}+\cdots+r_{i_1,t+v_1-1})+\cdots+a_{ji_n}(r_{i_nt}+\cdots+r_{i_n,t+v_n-1}) \quad \text{for } t = L+1,\ldots,T-L.$$

For notational convenience, we assume a common lead time L for all items, and $v_1, \ldots, v_n$ are the lot sizes for items $i_1, \ldots, i_n$, respectively. In the case when the lead times are different for each item, L will be indexed by i. The following result highlights the important effects of component commonality in an assembly supply chain.

PROPOSITION 6.13 *Suppose supply chain C_1 has a simple assembly structure and supply chain C_2 has a general assembly structure, and in C_2 the production of each component j is used to satisfy the demand for n items $i_1, \ldots, i_n$. If supply chains C_1 and C_2 both consist of component j, then*

(1) if item demands are i.i.d. random variables, the internal demand for component j in C_2 has a variance n times larger than that of C_1

(2) if item demands are correlated random variables with coefficient ρ and the same variances, then the internal demandfor component j in C_2 has a variance $n(1+n\rho-\rho)$ times that of C_1'swhen $\rho \geq \frac{1}{1-n}$.

The conventional wisdom in the automotive industry suggests that a high level of component sharing (a large n) combined with negatively correlated upper tier demand leads to smaller variation for lower tier demand. The proposition suggests that the demand variance for C_2 does decrease when the coefficient of correlation ρ is in between $-\frac{1}{n-1}$ and $-\frac{1}{n}$. However, when $-\frac{1}{n} < \rho \leq 1$ the variance of C_2 is greater than that of C_1. Further, as the the level of component sharing increases (as n increases), the coefficient range $\left[-\frac{1}{n-1}, -\frac{1}{n}\right]$ decreases by $\frac{1}{n(n-1)}$. In general, it seems more likely that component sharing leads to an increased variance for component demand. When the upper tier demands are perfectly correlated, the component demand variance increases by n^2. Nevertheless, if we consider the effects of component sharing on demand amplification, it does agree with the conventional wisdom, as stated in the following corollary.

COROLLARY 6.14 *Under the correlated case described in Proposition 6.11, the item-component demand amplification DA^s_{ij} between items $i_1, \ldots,$ i_n and component j changes by a factor of $\left(\sqrt{\frac{1+n\rho-\rho}{n}} - 1\right)$.*

Note that the only case when the factor $\left(\sqrt{\frac{1+n\rho-\rho}{n}} - 1\right) \geq 0$ is when $\rho = 1$. In all other cases, the *CV* at the component level is lower than the *CV* at the item level. Moreover, the degree of demand correlation as characterized by ρ has a more significant effect than the level of component sharing, as denoted by the value of n.

3.4 Capacity Levels

Setting capacity levels is a vital decision in all enterprises, and is certainly true in capital-intensive industries like automotive manufacturing. Here, we explore the effects of capacity to demand amplification independent of other effects in the supply chain.

Consider the case where all resource capacities are used to produce a particular item, and the capacity is preset at a constant level for all periods without dynamic changes. Consider three possible capacity settings: *maximum capacity* Cap_{max}, *average capacity* Cap_{avg}, and *actual capacity* Cap_{act}. Cap_{max} is set at the maximum period demand in the planning horizon. Cap_{avg} is set at the average demand, computed as the total demand in the planning horizon divided by the number of periods. Note that the average is the *minimum* "workable" capacity level, as any point less than this will be insufficient to maintain production. The actual capacity is bounded by the average and maximum levels, $Cap_{max} \geq Cap_{act} \geq Cap_{avg}$. Given this terminology, we define the *capacity smoothing coefficient* ε as:

$$\varepsilon = \frac{Cap_{max} - Cap_{act}}{Cap_{max} - Cap_{avg}}.$$

When $\varepsilon = 1$, the actual capacity is set at average demand. This implies that the volume of production is identical across all periods, and in each period the production consumes all resource capacity, resulting in a complete smoothing of production over all periods. Note that this 100% utilization of the capacity may not be desirable since discrete lot sizes would have to be broken apart. Conversely, when $\varepsilon = 0$, the actual capacity is set at the maximum demand level. This represents the cases where capacity is not binding. Under this condition, there is no smoothing due to capacity limitations. When $0 < \varepsilon < 1$, the actual capacity is between the maximum and the minimum values. Furthermore, the closer the actual capacity is to the minimum, the closer ε is to 1, and the greater the smoothing effect. This smoothing effect influences demand amplification in an assembly supply chain as described in the following result.

PROPOSITION 6.15 *Suppose component j of item i is manufactured in a facility, which has a capacity smoothing coefficient ε. ε has the following effect on item-component demand amplification (DA^{s}_{ij}):*

(1) when $\varepsilon = 0$, there is no demand amplification

(2) when $\varepsilon = 1$, the demand amplification is –CV(D)

(3) when $0 < \varepsilon < 1$, the demand amplification changes by $(\sqrt{c} - 1)$, where $c = \frac{Var(X)}{Var(D)} < 1$.

So, when resource capacity is binding $(0 < \varepsilon \leq 1)$, the demand emanating from a facility will have no more variation than the demand entering the facility and may actually be less.

Considering demand amplification in a multi-item environment, however, is more complex. Binding capacity constraints still limit the lot-sizes, but the capacity levels relative to the average demand for each item are not so tight because the capacity is established relative to the total demand for all items at that facility.

COROLLARY 6.16 *When overtime or outsourcing are used to increase capacity, the effect of capacity smoothing reduces, i.e., the capacity smoothing coefficient ε reduces, and the item-component demand amplification increases.*

Since overtime and outsourcing shift Cap_{act} closer to Cap_{max}, ε is reduced. The intuition behind this is simply that when more capacity options are available at the downstream facilities of the supply chain, a higher level of demand amplification should be expected.

4. Empirical Investigation

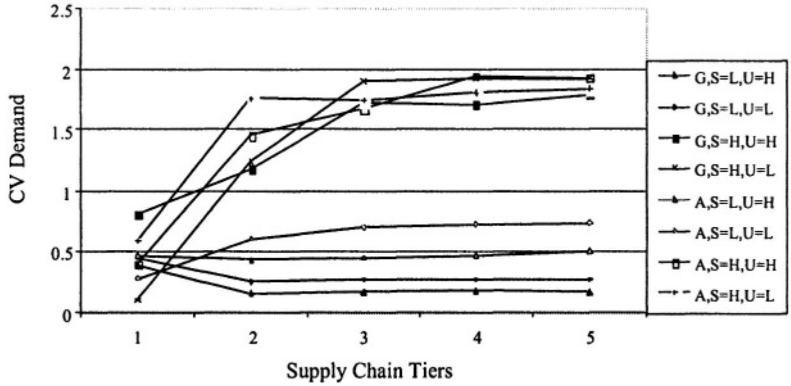
The above theoretical results are constructed for each factor separately, between a customer-supplier pair or two adjacent supply tiers. Although this is helpful for understanding the primary effects in assembly supply chains, considering more complex and realistic supply chain settings using empirical methods augments the theoretical insights. Accordingly, we have conducted a set of computational experiments to evaluate the joint effects of multiple factors in multiple supply tiers. This allows us to interpret the theoretical results in a broader context of assembly supply chains (Meixell and Wu, 1998; Meixell, 1998). In this section, we present the results of these experiments and interpret them in the context of the theoretical findings above.

4.1 Computational experiments

Data that represents a wide variety of supply chain environments are necessary to study behavior of manufacturing supply chains. Simulating a single instance of an actual operation has value in terms of developing an understanding of demand behavior, but would not provide the data needed for the factorial experimental design that tests for influence of a variety of factors on supply chain performance. For this reason, we construct a set of test instances by modifying the multi-item, multi-stage data set developed by Tempelmeier and Derstroff, 1996, using a supply chain with 40 items produced at 6 different facilities, distributed across 5 supply tiers with over 16 periods of demand. This dataset contains 1200 randomly generated problem instances that vary systematically in product and operation structure, setup time, time between order profile, design capacity, and setup cost. For each required treatment combination, the instances needed for this study were selected at random to fit the experimental design. In all cases, lead-time is set to one period and the initial inventory is 0.

We use analysis of variance and multiple regression to investigate the relationship between the experimental factors and the response variables. A detailed description of the experimental design and statistical analysis is available in Meixell and Wu, 1998. The experimental factors are the policies or costs pertaining to (1) order batching, (2) multiple releases, (3) component commonality, and (4) target capacity. The response variables are *total demand amplification (TDA)* and *total supply chain inventory.* TDA is computed as the *maximum* difference in demand variation between any two of the second through fifth supply tiers. The first tier demand is excluded from the calculation because the variation in end-item demand will bias the results while being exogenous to the system. *Total supply chain inventory* is calculated as the total inventory carried between periods, between the second through fifth tiers. The response variables are chosen for two purposes: to relate the empirical results to theoretical insights, and to understand broader impact of the design factors within the chain.

4.1.1 The effects of design factors in a dependent demand supply chain. The first experiment examines main factors that characterize an assembly supply chain. These include setup costs (high vs. low setup costs, which dictates the degree of order batching), component commonality (simple assembly vs. general assembly structure), targeted capacity utilization (90% vs. 50% utilization), and the coefficient of variation, *CV,* of end item demand (0.9 vs. 0.1). Eight replications are collected for each treatment combination. We analyze the main effects and interactions of the factors, and estimate the size of the differences, with the goal of identifying the factors that explain the differences in the response variables. Setup costs, component commonality, and capacity are all found to be significant to the demand amplification measure, TDA, at the 95% confidence level.



Component commonality: A (Simple Assembly Structure) G (General Assembly Structure)
Order batching: L (under low setup costs), H (under high setup costs)
Capacity utilization: H (90% utilization) L (50% utilization)

Figure 6.2. The effects of demand amplification by component commonality, order batching (setup costs), and capacity.

Figure 6.2 provides an overview of the experimental results relative to TDA. Setup costs appear to be the most pronounced of all factors: the two clusters of data points from tiers 2 onward are distinguished by high or low setup costs, where the low-setup-cost cluster has a mean TDA of 0.5, compared to 1.7 for the high-setup cluster. High setup costs, which encourage order batching, appears to drive more variation as demand propagates through the supply tiers regardless of component commonality and capacity utilization. Component commonality is another factor that is significant to TDA. As shown in Figure 6.2, products with *more* common components experience less demand variation in terms of CV in most cases. However, with high setup costs in lower supply tiers, we see that the opposite is true. This is due to the fact that higher setup cost encourages order batching at the upper tiers, which dominates the demand behavior at the lower tiers. Another notable interaction between setup costs and assembly structure is observed at the low setup costs cases. For general assembly structure, a slight demand de-amplification is observed – the *CV* is less at lower tiers than upper tiers in the supply chain. This behavior does not occur in the simple assembly structure.

The results also indicate that capacity utilization is a significant factor in demand amplification. More demand amplification occurs in supply chains at low utilization than at high utilization. Figure 6.2 also illustrates this effect. Note that for both simple assembly and general assembly structures, the high utilization line falls below the low utilization line. It appears that the variation in end-item demand is quickly "absorbed" by order batching *and* other forms of aggregation across supply tiers.

4.1.2 The effects of schedule release policies.

In the second experiment we investigate the influence of multiple schedule release policies on the response variables. In particular, we consider two types of schedule release policies similar to that considered in Propositions 6.11 and 6.12. The first is the commonly used freeze-up-to (FZ) policy, where the supplier uses the "most recent" schedule release from his immediate customer for demand calculation. The second policy uses the expected quantity (EQ) of customer schedule.

We first consider the effect of multiple schedule releases along with assembly structure on demand amplification. We find that schedule release policy does *not* significantly influence TDA at the 90% confidence level, and so we cannot reject the null hypothesis that the mean values of the treatment levels are equal. We then consider setup costs along with multiple schedule releases, and again, the release policy does not show significant influence when measured at the 90% confidence level, i.e. the two schedule release policies FZ and EQ do not appear to have a significant impact to TDA. On the other hand, when we examine the effects of the release policy to the second response variable, supply chain inventory, we found that the release policy is much more significant for this response (Table 6.1). Using the mean (EQ) instead of the latest customer release (FZ) appears to drive better inventory performance. In fact, we can reject the null hypothesis with 99% confidence that the means for inventory across the two schedule release policies are the same.

Response	**TDA**	**SC Inventory**
Source	Sum of Squares	Sum of Squares
MAIN EFFECTS		
A: Release policy	0.0002	140.8
B: Assembly structure	0.0614	902.0
INTERACTIONS		
AB	0.0237	6.156
RESIDUAL	1.106	2284.1
TOTAL (CORRECTED)	1.19092	3333.07

Table 6.1. ANOVA results when TDA or supply chain inventory is the response.

5. Comparing the Analytical and Empirical Results

The above highlights of empirical findings provide for further interpretation of the theoretical results, a complete description of the computational study can be found in Meixell and Wu, 1998). In this following, we examine the main results of the experiments in light of the theoretical predictions of Propositions 6.5-6.15.

On the effects of order batching and setup costs.
In the experiments, the effects of order batching are characterized by setup costs, as related to the inventory holding cost. The results as illustrated in Figure 6.2 confirm the basic intuition on order batching: in the low setup cost cases where less batching occurs, and little demand amplification is observed (Proposition 6.5), on the other hand, when setup costs are high, orders *are* consolidated into production batches, and the mean and variance for lower tier components will both increase (Proposition 6.6 and Corollary 6.7). In this case, more inventories are carried in the chain because of the larger variance in internal demand relative to the demand for the end item. While the *item-component demand amplification (Definitions 6.1 and 6.2)* measures the change in variation for a single item-component pair, the *tier-to-tier demand amplification (Definition 6.3)* measures the change in variation from one supply tier to another over all items. When the latter is used as a measure, Propositions 6.9 and 6.10 predict that demand amplification increases from one supply tier to another as a function of the lot size. Proposition 6.15 predicts that the setup capability of the system and system capacity will limit batch size, and therefore the amount of amplification. The experimental results again support these predictions. Not only does the tier-to-tier demand amplification increase after first batching occurs from tier 1 to tier 2, the amplification tapers off at the third tier since limited capacity prevents larger lot sizes from being formed.

On the effects of assembly structure and component commonality.
Proposition 6.13 and Corollary 6.14 state that when the components manufactured in the supply chain are shared by many downstream products, the fluctuation in end-item demand tends to create a greater level of variance. However, when put into the perspective of the mean, the amplification effects in terms of the change of *CV* for the general assembly structure are less than that of the simple assembly structure. The experimental results in Figure 6.2 support this insight, showing

that general assembly structures tend to experience less amplification than simple assembly structures. However, when order batching is not a factor (in the low-setup-cost cases), the general assembly structure could experience a demand de-amplification. This is again supported by Corollary 6.14, which predicts a negative demand amplification for general assembly structures when the upper tier demands are independent. When the components manufactured in the supply chain are shared by a large number of downstream products, the fluctuation in end-item demands tends to create a more significant level of variance. When the product demands are positively correlated, this variance will be further exaggerated. Conversely, when the demands are negatively correlated, the variance decreases by comparison.

This result provides some insight for redesign or consolidation of existing products in the supply chain to increase the commonality of components. If the upper tier demands tend to be positively correlated, as is the case for a class of components in the automotive industry, the demand variance will amplify through the supply tiers, prompting an unreasonably high level of reserved capacity in upstream facilities.

On the effect of capacity and utilization.
Propostion 6.15 states that when the effects of order batching are excluded from consideration, limited capacity has a smoothing effect on demand amplification. In other words, when manufactured in a capacity-bounded facility, the orders emanating from the facility tend to have a lower variance than the orders entering it. As pointed out above, this phenomenon explains the tendency of a tapered demand amplification at lower supply tiers as exhibited in Figure 6.2. Since the test problems used in the experiments have a capacity level close to the true expected demand, the observed amplification profile is well explained by the proposition.

Also, high capacity utilization drives more inventories in a supply chain, regardless of the assembly structure. This is intuitive as limited capacity would prompt some items to be built ahead for demands realized in a later period, it follows that higher capacity utilization *drives higher supply chain inventory.* The capacity smoothing phenomenon explains a regular, but artificial, smooth order pattern upstream in a dependent demand supply chain, despite of the fluctuation in end-item demand. Contrary to conventional wisdom in the industry, this result suggests that in reality the supply chain behaves in a more stable and consistent manner over time when its capacity is close to its true expected demand, but a higher inventory level is to be expected. Setting capacity based on peak demand or artificially restricting inventory levels is likely to create a higher level of variance in the supply chain. The capacity issues become more complex when strict end item due dates are imposed and expediting and outsourcing activities come into play.

On the effect of multiple schedule release policies.
As stated in Proposition 6.11, if the schedule releases and the period-to-period demands are i.i.d., demand amplification across multiple releases does not differ from single release. However, when each schedule release is dependent on the previous release, Proposition 6.12 suggests that it is preferable to follow the *expected* product quantity over up-to-date releases rather than following the most recent release. The experiment shows that while no significant difference is observed in *demand amplification* between the two release policies, the difference in *supply chain inventory* is significant (see Table 6.1). This corresponds to the prediction that a higher variance is to be expected for the FZ policy, which results in a higher inventory level. Multiple schedule releases is a common source of frustration for production managers in automotive supply chains where a release may be changed several times before final shipment. As suggested by our study, it is advisable to keep track of recent releases and follow the expected quantity rather than the latest release. In essence, this result suggests that it is possible to reduce the variance created by multiple schedule releases by focusing on the underlying *population* of schedules that could be released, rather than any single realization. Such populations may be estimated using an empirical distribution, perhaps using the scheme described in Proposition 6.12, or by using historic order data to construct *a priori* distributions.

6. Conclusions

Demand propagation is a fundamental behavior of supply chains – a deeper understanding of this phenomenon is essential to the overall improvement of performance. Specifically, we find four main factors driving demand propagation in assembly supply chains: order batching, component commonality, production capacity, and multiple schedule releases. We examine theoretical guidelines on the demand behavior that may provide managerial insights, and a few highlights of empirical results are given which put the theoretical insight in practical contexts. In the following, we summarize main findings of the research that may suggest answers to important managerial questions.

Which of the supply chain design factors has the greatest influence on overall performance?
Setup cost appears to have the most significant influence on the perfor-

mance of assembly supply chains. In specific, high setup costs tend to increase batch size, which increase demand variation along the supply tiers. Assembly structure also appears to have a significant influence. While a higher degree of component sharing may lead to higher demand variance for a particular component, it tend to drive down demand variation *(CV)* in the supply tier. Utilization is another source of influence. Binding capacity in any period at any facility reduces the effects of demand amplification, which also reduces downstream demand variation. The variation in end-item demands, however, does not appear to have a significant impact to overall demand behavior.

What combination of supply chain design factors lead to better performance?

Designing a supply chain with a high level of component sharing (general assembly structure), low setup costs, and high capacity utilization throughout the chain can reduce or eliminate demand amplification. These represent the most favorable conditions for supply chain performance if the intention is to reduce demand amplification. On the other hand, low setup costs and low utilization result in the least amount of inventory in the supply chain, regardless of the assembly structure. Some inventory must be incurred, then, if utilization is high, even when setup costs are low. Inventory and demand amplification do not have identical response to the main design factors – low utilization improves inventory performance, and high utilization improves demand amplification.

In the typical ERP environment of multiple schedule releases, does frequent schedule updates influence supply chain performance? If so, how to overcome this problem?

The practice of using the most recent scheduling release from upper tier customers have a negative impact to supply chain inventory. Our study find that it is better to treat scheduled quantity as a random variable, and schedule production using the expected value of that demand. While this does not appear to have a significant impact on demand amplification, it does have a significant impact on the inventory performance (55% inventory reduction in our experiments). Since the inventory performance often relates to demand uncertainty, using expected demand values help to reduce demand uncertainty across supply tiers, and thereby reducing the needs for keeping high inventory levels.

Acknowledgements

This research is supported, in part, by National Science Foundation Grant DMI-9634808, Ford Motor Company, and Agere Systems (formerly the Microelectronics Division of Lucent Technologies).

Appendix

Proof of Proposition 6.6. Denote $x_{j1}, x_{j,v+1}, \ldots, x_{j,T-v+1}$ the realization of random variables $X_1, \ldots, X_{T-v+1}$, and $r_{i1}, r_{i2}, \ldots, r_T$ that of random variables $D_1, \ldots, D_T$, respectively, where $x_{j,t-L_i} = a_{ji}(r_{it} + r_{i,t+1} + \cdots + r_{i,t+v-1})$, for $t = L_i + 1, \ldots, T - L_i$. $D_1, D_2, \ldots, D_T$ are i.i.d. random variables with expected value *E(D)* and variance *Var(D)*. As a result of lot-sizing, v periods of orders for component j are combined to satisfy the internal demands imposed by item i, resulting in a stream $X_1, X_{v+1}, \ldots, X_{T-v+1}$ for non-zero production periods where $X_1 = a_{ji}(D_1 + D_2 + \cdots + D_v)$, $X_{v+1} = a_{ji}(D_{v+1} + \cdots + D_{2v})$, …, $X_{T-v+1} = a_{ji}(D_{T-v+1+} + \cdots + D_T)$. The distribution of *X* is the convolution of that of D_t's. Since D_t's are i.i.d. random variables, we have

$$E(X) = a_{ji} v E(D) \text{ and } Var(X) = a_{ji}^2 v \, Var(D).$$

This confirms statement (1).

Now, the item-component demand amplification is

$$\begin{aligned} DA_{ij}^s &= CV(X) - CV(D) \\ &= \frac{\sqrt{a_{ji}^2 \cdot v \cdot Var(D)}}{a_{ji} \cdot v \cdot E(D)} - \frac{\sqrt{Var(D)}}{E(D)} \\ &= \left(\frac{1}{\sqrt{v}} - 1\right) \cdot CV(D). \end{aligned}$$

This confirms statement (2). □

Proof of Corollary 6.7. The proof is similar to that of Proposition 6.6 except that now $E(X) = a_{ji} v E(D)$, but

$$\begin{aligned} Var(X) &= Var\left(a_{ji} D_1 + a_{ji} D_2 + \cdots + a_{ji} D_v\right) \\ &= \sum_{t=1}^{v} a_{ji}^2 \cdot Var(D_t) + 2 \sum_{1 \le t < s \le v} a_{ji}^2 \cdot Cov(D_t, D_s). \end{aligned}$$

If $D_1, D_2, \ldots, D_T$ have the same variance, i.e., $Var(D_t) = Var(D)$, then $Cov(D_t, D_s) = \rho \, Var(D)$ and $Var(X) = a_{ji}^2 \cdot v \cdot Var(D) + a_{ji}^2 (v^2 - v) \cdot \rho \cdot Var(D) = a_{ji}^2 v(1 + v\rho - \rho) \, Var(D)$, we have

$$\begin{aligned} DA_{ij}^s &= CV(X) - CV(D) \\ &= \frac{\sqrt{a_{ji}^2 \cdot v(1 + v\rho - \rho) \cdot Var(D)}}{a_{ji} \cdot v \cdot E(D)} - \frac{\sqrt{Var(D)}}{E(D)} \\ &= \left(\frac{\sqrt{v(1 + v\rho - \rho)}}{v} - 1\right) \cdot CV(D). \end{aligned}$$

This confirms statement (1). □

Statement (2) follows by setting $\rho = 1$.

Proof of Lemma 6.8. Without loss of generality, let $a_{ji} = 1$, denote $\mu = E(Y) = E(D)$

$$\begin{aligned} Y &= \begin{cases} a_{ji}E(D_1 + \cdots + D_v) & \text{with probability } \frac{1}{v} \\ 0 & \text{with probability } 1 - \frac{1}{v} \end{cases} \\ Var(Y) &= \frac{v-1}{v}(0-\mu)^2 + \frac{1}{v}(v \cdot \mu - \mu)^2 = \frac{(v-1)\mu^2}{v}(1 + v - 1) = (v-1)\mu^2 \\ CV(Y) &= \frac{\sqrt{(v-1)\mu^2}}{\mu} = \sqrt{v-1} \end{aligned}$$

Proof of Proposition 6.9. Since the demands are drawn from i.i.d. random variables $D_1, \ldots, D_T$, we know that $Var(Y) = a_{ji}^2 v\, Var(D)$. Without loss of generality, let $a_{ji} = 1$, denote $\mu = E(Y) = E(D)$. Then, from Lemma 6.8 we have $Var(Y) = (v-1)\mu^2 = v \cdot Var(D)$. Thus, $Var(D) = \frac{(v-1)\mu^2}{v}$, and $CV(D) = \sqrt{\frac{v-1}{v}}$. Therefore, the tier-to-tier demand amplification is

$$\begin{aligned} TA_{kl} &= CV(Y) - CV(D) \\ &= \sqrt{v-1} - \sqrt{\frac{v-1}{v}} \\ &= \frac{(v-1)\sqrt{(v-1)}}{v} \end{aligned}$$

which is an increasing function of the lot size $v > 0$. □

Proof of Proposition 6.10. Since $Y1, \ldots, YN$ are i.i.d. random variables, by applying Lemma 6.8 the variance of production orders $(Y1 + \cdots + YN)$ is

$$\begin{aligned} Var(Y1 + \cdots + YN) &= Var(Y1) + \cdots + Var(YN) \\ &= N(v-1)\mu^2. \end{aligned}$$

This confirms statement (1).

$$CV(Y1 + \cdots + YN) = \frac{\sqrt{N(v-1)\mu^2}}{N\mu} = \sqrt{\frac{v-1}{N}}.$$

This confirms statement (2).

$$\begin{aligned} Var(D1 + \cdots + DN) &= N \cdot \frac{(v-1)\mu^2}{v} \\ CV(D1 + \cdots + DN) &= \sqrt{\frac{v-1}{vN}} \\ TA_{kl} &= CV(Y1 + \cdots + YN) - CV(D1 + \cdots + DN) \\ &= \sqrt{\frac{v-1}{N}} - \sqrt{\frac{v-1}{vN}} \end{aligned}$$

$$= \sqrt{\frac{v-1}{N}}\left(1-\frac{1}{\sqrt{v}}\right).$$

The above term is an increasing function of the lot size v, while a decreasing function of *N*, the number of components under production. □

Proof of Proposition 6.11. Denote $x_{j1}^{1}, x_{j,v+1}^{1}, \ldots, x_{j,T-v+1}^{1}, \ldots, x_{j1}^{\Theta}, \ldots, x_{j,T-v+1}^{\Theta}$ the realization of random variables $X_1^1, X_{v+1}^1, \ldots, X_{T-v+1}^1, \ldots, X_1^{\Theta}, \ldots, X_{T-v+1}^{\Theta}$, and $r_{i1}^1, \ldots, r_{iT}^1, \ldots, r_{i1}^{\Theta}, \ldots, r_{iT}^{\Theta}$ that of random variables $D_1^1, \ldots, D_T^1, \ldots, D_1^{\Theta}, \ldots, D_T^{\Theta}$, respectively, where $x_{j,t-L_i} = a_{ji}(r_{it} + r_{i,t+1} + \cdots + r_{i,t+v-1}$, for $t = L_i + 1, \ldots, T - L_i$. Suppose $D_1^1, \ldots, D_T^1, \ldots, D_1^{\Theta}, \ldots, D_T^{\Theta}$ are i.i.d. random variables with expected value *E(D)* and variance *Var(D)*. Similar to Proposition 6.6, we have

$$E(X) = vE(D) \text{ and } Var(X) = v\,Var(D).$$

Therefore

$$\begin{aligned} DA_{ijt}^{m} &= CV(x_{j,t-L_i}^{k+1,\tau}, \tau = 1, \ldots, \Theta, t = 1, \ldots, T - L_i) - \\ &\qquad CV(r_{it}^{k,\tau}, \tau = 1, \ldots, \Theta, t = L_i + 1, \ldots, T) \\ &= CV(X) - CV(D) \\ &= \left(\frac{1}{\sqrt{v}} - 1\right) \cdot CV(D). \end{aligned}$$

□

Proof of Proposition 6.12. For the freeze-up-to schedule, define the first period $X_1^1 = S_1$, and define the rest of the period by the vector-valued random variable

$$S = \begin{bmatrix} X_{v+1}^2 \\ X_{2v+1}^3 \\ \vdots \\ X_{T-v+1}^{\Theta} \end{bmatrix} = \begin{bmatrix} S_2 \\ S_3 \\ \vdots \\ S_{T/v} \end{bmatrix}.$$

The mean and variance of *S* are as follows:

$$E(S) = \begin{bmatrix} E(X_{v+1}^2) \\ E(X_{2v+1}^3) \\ \vdots \\ E(X_{T-v+1}^{\Theta}) \end{bmatrix} = E(\bar{X}).$$

Var(S) is a matrix $\Sigma_s = [\sigma_{ij}]$ where each entry $\sigma_{ij} = Cov(S_i, S_j)$. But, $Cov(S_i, S_j) = 0$ for $i \neq j$ and $Cov(S_i, S_j) = Var(S_i)$ for $i = j$. Therefore, Σ_s is a scalar matrix and $Var(S) = \Sigma_s = Var(X) \cdot I$. Similarly, the mean and variance of random vector $\bar{X}$ are as follows:

$$E(\bar{X}) \;=\; E(S),\, Var(\bar{X}_1) = Var(S_1),$$

for the first period. And from the $(v + 1)^{st}$ period on:

$$Var(\bar{X}) \;=\; \sigma_X = \begin{bmatrix} Var(\bar{X}_{v+1}) \\ Var(\bar{X}_{2v+1}) \\ \vdots \\ Var(\bar{X}_{T-v+1}) \end{bmatrix}^{T} \cdot I = \begin{bmatrix} Var(X)/v \\ Var(X)/(2v) \\ \vdots \\ Var(X)/((T/v)v) \end{bmatrix}^{T} \cdot I$$

$$= \frac{Var(X)}{v} \cdot \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ v/T \end{bmatrix}^{\mathrm{T}} \cdot I = \frac{Var(X)}{v} \cdot c \cdot I$$

where c is a fixed vector. Therefore, $Var(\bar{X}) = Var(S) \cdot c/v$. □

Proof of Proposition 6.13. We first distinguish the simple and the general assembly structures as follows. For component j and item i in a *simple* assembly structure, define g_0 as the realization of random variable G_0. Then,

$$x_{j,t-L_i} = g_0 \text{ where } g_0 = a_{ji}(r_{it} + r_{i,t+1} + \cdots + r_{i,t+v-1}).$$

For component j and items $i_1, \ldots, i_n$ in a *general* assembly structure, define $z_{j1}, z_{j,v+1}, \ldots, z_{j,T-v+1}$ as the realization of random variables $Z_1, \ldots, Z_{T-v+1}$ and $g_1, \ldots, g_n$ as the realization of random variables $G_1, \ldots, G_n$. Then,

$$z_{j,t-L} = g_1 + \cdots + g_n \text{ where } g_k = a_{ji_k}(r_{i_k t} + \cdots + r_{i_k,t+v_k-1}).$$

Thus, we have $E(Z) = \sum_{k=1}^{n} E(G_k)$ and

$$Var(Z) = Var(G_1 + G_2 + \cdots + G_n) = \sum_{k=1}^{n} Var(G_k) + 2 \sum_{1 \le k < l \le n} Cov(G_k, G_l).$$

If $G_0, G_1, \ldots, G_T$ are i.i.d. random variables, then $Cov(G_k, G_l) = 0$ and $Var(G_k) = Var(G_0)$. Therefore, $Var(Z) = n\,Var(G_k) = n\,Var(G_0)$. This confirms statement (1). If $G_1, G_2, \ldots, G_T$ are correlated, since

$$\begin{aligned} Var(G_k) &= Var(G_l) = Var(G_0) \qquad \forall k, l \\ Var(Z) &= n \cdot Var(G_0) + 2 \sum_{1 \le k < l \le n} \rho \sqrt{Var(G_k) \cdot Var(G_l)} \\ &= n \cdot Var(G_0) + (n^2 - n) \cdot \rho \cdot Var(G_0) \\ &= n(1 + n\rho - \rho)\,Var(G_0). \end{aligned}$$

This corresponds to statement(2). But since by definition,

$$Var(Z) = n(1 + n\rho - \rho)\,Var(G_0) \ge 0$$

it must be that $\rho \ge \frac{1}{1-n}$. □

Proof of Corollary 6.14. The item-component demand amplification

$$\begin{aligned} DA_{ij}^{s} &= CV(Z) - CV(G_k) \\ &= \frac{\sqrt{Var(Z)}}{\sum_{k=1}^{n} E(G_k)} - \frac{\sqrt{Var(G_k)}}{\frac{1}{n}\sum_{k=1}^{n} E(G_k)}. \end{aligned}$$

From Proposition 6.13, we have $Var(Z) = n(1 + n\rho - \rho)\,Var(G_0)$, and $CV(G_k) = CV(G_0)\ \forall k$ therefore

$$DA_{ij}^{s} = \frac{\frac{1}{n}\sqrt{n(1 + n\rho - \rho)\,Var(G_k)}}{\frac{1}{n}\sum_{k=1}^{n} E(G_k)} - \frac{\sqrt{Var(G_k)}}{\frac{1}{n}\sum_{k=1}^{n} E(G_k)}$$

$$= \left(\sqrt{\frac{(1+n\rho-\rho)}{n}}-1\right)CV(G_k)$$
$$= \left(\sqrt{\frac{(1+n\rho-\rho)}{n}}-1\right)CV(G_0).$$

□

Proof of Proposition 6.15. When $\varepsilon = 0$ capacity is non-binding, thus, from Proposition 6.5 there is no demand amplification $(DA^s_{ij} = 0)$. This confirms statement (1).

Using the notation similar to before and denoting random variables $D_1, \ldots, D_T$ as the demand for item i, and random variables$X_1, \ldots, X_T$ as the production quantity for component j after production has been scheduled. Denote $x_{j1}, x_{j2}, \ldots, x_{jT}$, as the occurrences of random variables $X_1, \ldots, X_T$, and $r_{i1}, r_{i2}, \ldots, r_{iT}$ that of $D_1, \ldots, D_T$. For simplicity, assume $a_{ji} = 1$.

When $\varepsilon = 1$, the production quantity for each period must be set equal to the capacity Cap_{act} which is set to Cap_{avg}, i.e., $x_{j1} = x_{j2} = \cdots = x_{jT} = \frac{1}{T}\sum_{t=1}^{T} r_{it} = Cap_{avg} = Cap_{act}$. Thus, $E(X) = Cap_{act} = E(D)$ and *Var(X)* = 0. Therefore,

$$DA^s_{ij} = CV(X) - CV(D) = 0 - \frac{\sqrt{Var(D)}}{E(D)} = -CV(D).$$

This confirms statement (2).

When $0 < \varepsilon < 1$, the capacity is binding for at least one demand period, i.e. $\exists r_{ik}$, $(r_{ik} - Cap_{act}) = \xi_{ik} > 0$. As a result, the demand r_{ik} will be produced in two parts, $(r_{ik} - \xi_{ik})$ and ξ_{ik}. Or, more precisely, demand r_{ik} will be satisfied by at least two productions $x_{jl} \le (r_{ik} - \xi_{ik})$ and $x_{jm} \ge \xi_{ik}$ and $x_{jl} + x_{jm} = r_{ik}$. To satisfy demand $r_{i1}, r_{i2}, \ldots, r_{iT}$ with production $x_{j1}, x_{j2}, \ldots, x_{jT}$ we need to deal with two subsets of demands, say *K* and *L:* for $k \in K$, we have $(r_{ik} - Cap_{act}) = \xi_{ik} > 0$ and for $l \in L$, we have $(r_{il} - Cap_{act}) = \xi_{il} \le 0$. In other words, we use excess capacity $\left(\sum_{l\in L}\xi_{il}\right)$ in some periods to cover excess demands $\left(\sum_{k\in K}\xi_{ik}\right)$ in others. It is then easy to see that the variance for production will be smaller than the variance for demands, i.e., *Var(X)* < *Var(D)* if set *K* is nonempty. Set $c = Var(X)/Var(D) < 1$, then we have

$$DA^s_{ij} = CV(X) - CV(D) = \frac{\sqrt{c \cdot Var(D)}}{E(D)} - \frac{\sqrt{Var(D)}}{E(D)} = (\sqrt{c} - 1) \cdot CV(D).$$

This confirms statement (3). □

References

Arntzen, B.C., Brown, G.G., Harrison, T.P., and Trafton, L.L. (1995). Global supply chain management at Digital Equipment Corporation. *Interfaces,* 25(1):69–93.

Bahl, H.C., Ritzman, L.P., and Gupta, J.N.D. (1987). Deterministic lot sizes and resource requirements: A review. *Operations Research,* 35:329–345.

Balakrishnan, A. and Geunes, J. (2000). Requirements planning with substitutions: Exploiting bill-of-materials flexibility in production planning. *Manufacturing & Service Operations Management,* 2(2):166–185.

Bhatnagar, R., Chandra, P., and Goyal, S.K. (1993). Models for multiplant coordination. *European Journal of Operational Research,* 67:141–167.

Camm, J.D., Chorman, T.E., Dill, F.A., Evans, J.R., Sweeney, D.J., and Wegryn, G.W. (1997). Blending OR/MS, judgement, and GIS: Restructuring P&G's supply chain. *Interfaces,* 27(1):128–142.

Chen, F., Drezner, Z., Ryan, J.K., and Simchi-Levi, D. (1999). The bullwhip effect: Managerial insights on the impact of forecasting and information on variability in a supply chain. In Tayur, S., Ganeshan, R., and Magazine, M., editors, *Quantitative Models for Supply Chain Management.* Kluwer Academic Publishers, Boston, Massachusetts.

Ertogral, K. and Wu, S.D. (2000). Auction-theoretic coordination of production planning in the supply chain. *IIE Transactions on Design and Manufacturing,* 32(10):931–940. Special Issue on Decentralized Control of Manufacturing Systems.

Ganeshan, R., Jack, E., Magazine, M.J., and Stephens, P. (1999). A taxonomic review of supply chain management research, in quantitative models for supply chain management. In Tayur, S., Ganeshan, R., and Magazine, M., editors, *Quantitative Models for Supply Chain Management.* Kluwer Academic Publishers, Boston, Massachusetts.

Geoffrion, A.M. and Power, R.F. (1995). Twenty years of strategic distribution system design: An evolutionary perspective. *Interfaces,* 25:105–128.

Graves, S.C., Kletter, D.B., and Hetzel, W.B. (1998). Dynamic model for requirements planning with application to supply chain optimization. *Operations Research,* 46(3):S35–S49.

Karabuk, S. and Wu, S.D. (2002). Decentralizing semiconductor capacity planning via internal market coordination. *IIE Transactions on Operations Engineering,* 34:743–759. Special Issue in Advances in Large-Scale Optimization for Logistics, Production and Manufacturing Systems.

Karabuk, S. and Wu, S.D. (2003). Coordinating strategic capacity planning in the semiconductor industry. *Operations Research,* 51(6).

Kekre, S., Mukhopadhyay, T., and Srinivasan, K. (1999). Modeling impacts of electronic data interchange technology. In Tayur, S., Ganeshan, R., and Magazine, M., editors, *Quantitative Models for Supply Chain Management.* Kluwer Academic Publishers, Boston, Massachusetts.

Kimms, A. (1997). *Multi-Level Lot Sizing and Scheduling: Methods for Capacitated, Dynamic, and Deterministic Models.* Physica-Verlag, Heidelberg, Germany.

curement management. *Hewlett-Packard Journal,* pages 28–38.

Lederer, P.J. and Li, L. (1997). Pricing, production, scheduling, and delivery-time competition. *Operations Research,* 45:407–420.

Lee, H.L. (1996). Effective inventory and service management through product and process redesign. *Operations Research,* 44:151–159.

Lee, H.L. and Billington, C. (1993). Material management in decentralized supply chains. *Operations Research,* 41:835–847.

Lee, H.L. and Billington, C. (1995). The evolution of supply-chain-management models and practice at Hewlett-Packard. *Interfaces,* 25 (5):42–63.

Lee, H.L., Padmanabhan, V., and Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science,* 43:546–558.

Levy, D. (1997). Lean production in an international supply chain. *Sloan Management Review,* 38:94–102.

Maes, J. and Wassenhove, L.N. Van (1988). Multi-item singel-level capacitated dynamic lot-sizing heuristics: A general review. *Journal of the Operational Research Society,* 39:991–1004.

Martin, C.H., Dent, D.C., and Eckhart, J.C. (1993). Integrated production, distribution and inventory planning at Libbey-Owens-Ford. *Interfaces,* 23(3):68–78.

Meixell, M.J. (1998). *Modeling Demand Behavior in Manufacturing Supply Chains.* PhD thesis, Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, Pennsylvania.

Meixell, M.J. and Wu, S.D. (1998). Demand behavior in manufacturing supply chains: A computational study. ISE Technical Report 98T-008, Lehigh University, Bethlehem, Pennsylvania.

Nam, S. and Logendran, R. (1992). Aggregate production planning – a survey of models and methodologies. *European Journal of Operational Research,* 61:255–272.

O'Brien, C. and Head, M. (1995). Developing a full business environment to support just-in-time logistics. *International Journal of Production Economics,* 42:41–50.

Robinson, Jr., E.P., Gao, L.L., and Muggenborg, S.D. (1993). Designing an integrated distribution system at DowBrands, Inc. *Interfaces,* 23(3):107–117.

Singh, J. (1996). The importance of information flow within the supply chain. *Logistics Information Management,* 9:28–30.

Slats, P.A., Evers, B. Bhola J.J.M., and Dijkhuisen, G. (1995). Logistic chain modeling. *European Journal of Operational Research,* 87:1–20.

Srinivasan, K., Kekre, S., and Mukhopadhyay, T. (1994). Impact of electronic data exchange on JIT shipments. *Management Science,* 40(10):1291–1304.

Sterman, J.D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision-making experiment. *Management Science,* 35:321–339.

Takahashi, K.R., Muramatsu, R., and Ishii, K. (1987). Feedback methods of production ordering systems in multi-stage production and inventory systems. *International Journal of Production Research,* 25:925–941.

Tempelmeier, H. and Derstroff, M. (1996). A Lagrangean-based heuristic for dynamic multilevel multiitem constrained lotsizing with setup times. *Management Science,* 42:738–757.

Thomas, D.J. and Griffin, P.M. (1996). Coordinated supply chain management. *European Journal of Operational Research,* 94:1–15.

Towill, D.R. (1992). Supply chain dynamics – the change engineering challenge for the mid 1990's. *Proceedings of the Institute of Mechanical Engineers, Part B: Journal of Engineering Manufacture,* 206:233–245.

Veinott, Jr., A.F. (1969). Minimum concave-cost solution of Leontief substitution models of multi-facility inventory systems. *Operations Research,* 17:262–291.

Vidal, C.J. and Goetschalckx, M. (1997). Strategic production-distribution models: A critical review with emphasis on global supply chain models. *European Journal of Operational Research,* 98:1–18.

Wu, S.D. and Golbasi, H. (2003). Multi-item, multi-facility supply chain planning: Models, complexities, and algorithms. *Computational Optimization and Applications.* Under revision.

II

ELECTRONIC COMMERCE AND MARKETS

This page intentionally left blank

Chapter 7

BRIDGING THE TRUST GAP IN ELECTRONIC MARKETS: A STRATEGIC FRAMEWORK FOR EMPIRICAL STUDY

Gary E. Bolton, Elena Katok
Smeal College of Business
Penn State University

Axel Ockenfels
Department of Economics
University of Cologne
Cologne, Germany

Abstract Trust that suppliers and buyers will keep their word is a necessary ingredient to a well functioning marketplace. Nowhere is the issue trickier than for electronic markets, where transactions tend to be geographically diffuse and anonymous, putting them out of the reach of the legal safeguards and the long-term relationships that build trust in the brick-and-mortar world. Many online platforms have turned to automated reputation systems as a way of giving traders a heads-up on who they are dealing with. Here we describe a strategic framework for thinking about these systems. We also present some lab data that provides an initial sense of effectiveness. We find that reputation has substantial positive effect, but not enough to be a close substitute for personal relationships; this is so even though our laboratory test abstracts away from many of the problems reputation systems must confront in the field. The evidence also suggests directions for improving automated reputation system performance.

1. Introduction: The problem of trust in electronic markets

According to a recent report by the Aberdeen Group, the market for strategic e-sourcing tools is growing at a rate of 98% per year, significantly higher than the growth rate of the IT industry as a whole (Morphy, 2001). Most of the emphasis has been on creating ways for improving access to suppliers, facilitating competition, and driving down prices. This is the promise of the Internet: expanded, more efficient markets. But the promise is not without peril. Realizing gains from greater competition means moving away from the long-term relationships commonly used to safeguard trust in brick-and-mortar markets. Lack of trust could undo all the good of increased competition and then some (Arrow, 1974).

The potential and the pitfalls are evident in an illustrative case a group of insurance and B2B executives took up during a recent conference on trust in global B2B e-commerce: Sarah Jones works for the American clothing retailer Blazer Barn. She is searching the Internet for a new blazer manufacturer when she comes upon Hong Kong Blazer, posted on the B2B exchange Buyeverything.com. Hong Kong Blazer manufactures the style of blazer Sarah is looking for, at a good price, and offers delivery in six weeks, in time for the back-to-school season. The fact that Sarah has never heard of Hong Kong Blazer, let alone dealt with them, makes her nervous, but she decides to go ahead anyway. She completes the necessary contractual documents, some of which are in Chinese, a language she does not understand, and then waits with fingers crossed. Unfortunately things do not go well. The wrong style of blazer shows up several weeks late, the shipment short a thousand units. Hong Kong Blazer blames the shipping company, and anyway Sarah did not specify the order properly. Given the differences in American/China jurisdictions, Sarah has little legal recourse.[1]

The legal problems surrounding e-commerce are not limited to international trade. Within the U.S., e-commerce transactions often span multiple jurisdictions with conflicting laws, making legal recourse difficult if not impossible (Federal Bureau of Investigation, 2002). Some of these problems are transient. But even so, the law is unlikely to become a complete solution because, as in the brick-and-mortar economy, legal recourse will be expensive and time consuming.

Brick-and-mortar markets heavily supplement the law with informal trust building mechanisms, particularly long-term relationships, but also

[1] Dennehy (2000) reports on this conference, and participant's reaction to this case.

word-of-mouth reputation. Both are usefully thought of as forms of reciprocity. Long-term trading relationships are a form of *direct reciprocity;* I continue to deal with you so long as you were reliable with me in the past (Axelrod, 1984). Word-of-mouth reputation is a form of *indirect reciprocity;* I continue to deal with you so long as you were reliable with others in the past (Alexander, 1987). As noted, moving away from long-term relationships is the flip side of exploiting the scope of Internet markets. But one need not leave word-of-mouth behind: Several electronic trading platforms have introduced online "feedback" mechanisms intended to institutionalize word-of-mouth reputation by storing and distributing information about the transactions of those they mediate.

The question we explore in this chapter is whether online reputation systems can fill the trust gap, or at least that part that long-term relationships filled. It turns out that in theory, the answer is yes: direct and indirect reciprocity can be equally effective. As we will argue later, the latter may require more information to function properly, but information dissemination is precisely what the Internet excels at. That said, we also present laboratory evidence that the reality is more complicated: We find that even under conditions that are arguably ideal, reputation systems fail to match the performance of long-term relations, although importantly, reputation systems do a lot of good relative to not having anything at all. In addition, our experiments pinpoint some problems and point in directions to improve the systems.

Theory and experiment are compliments to field evidence, and it is important to interpret our results in the context of what is known from the field. In fact, formal reputation systems and markets have a long and successful history together. One particularly well-documented example dates to the beginning of the first millennium, involving a group of merchants known as the Maghribis, who operated and traded throughout the Mediterranean region. At the time, ship navigation was a hazardous business. A merchant generally required a business associate to sell (or buy) the goods and handle profits (or expense money) at the other end of the transaction. The Maghribis developed a very explicit system for disseminating information about which associates could be trusted and which could not. When a Maghribi trader discovered an associate was cheating him, he was under obligation to write a letter to all the other Maghribis warning them to stay away from this associate. This word-of-mouth system provided a powerful incentive for associates to mind their reputation less they lose business (Greif, 1989). It is essentially a one-way reputation system of the kind we investigate in Section 2.

The development of reputation systems was also an important factor in the expansion of trade that took place during the European Renaissance (Milgrom, North, and Weingast, 1990). At a deeper (evolutionary) level, research suggests that reputation mechanisms play a central role in all human social behavior, as a regulator of moral systems (Alexander, 1987). This idea has recently been analyzed in the form of an 'image scoring game,' a game we investigate in Section 3 to illustrate some of the issues in two-way reputation systems.

The new electronic feedback mechanisms essentially institutionalize the old word-of-mouth methods. eBay, for example, hosts a 'feedback forum,' enabling buyers and sellers to rate each other after the transaction. This information is then made available to all potential future transaction partners. Bizrate.com asks consumers to complete surveys on registered retailers, and then converts this information into store ratings. In the B2B arena, Reputation.com provides software that allows businesses to track and internally disseminate information about vendor performance across the organization. Other reputation mechanisms on the Internet include 'expert' evaluation of past performance such as expertcentral.com, and product review such as epinions.com. While diverse in form, all of these methods are based on the same fundamental principle of indirect reciprocity: the dissemination of information on individual past performance with others can be used to build trust and trustworthiness.[2]

The reliability and limits of these modern systems are only beginning to be investigated. Field studies of online auction platforms find that reputation mechanisms have at least some of the desired economic effect, though these studies do not present a completely coherent picture. Some find that reputable sellers are more likely to sell their items but with no price effect (e.g., Resnick and Zeckhauser, 2002), other studies find that negative feedback reduces the attainable price (e.g., Lucking-Reiley, Bryan, Prasad, and Reeves, 2000), and a few studies find that positive feedback induces price premiums (e.g., Houser and Wooders, 2000).[3]

Further field work should clear up some of the discrepancies. But there are also obstacles. The complexity of the natural environment makes it

[2] See `http://pages.ebay.com/services/forum/feedback.html` for eBay's feedback forum, similar to mechanisms used by Amazon and Yahoo, among others. For Bizrate's system, see `http://www.bizrate.com/content/about.xpml`. Traditional providers of reputation information, such as accounting firms or business credit rating companies, are also active on the Internet.

[3] Other field studies include Ba and Pavlou (2002), Dewan and Hsu (2001), Melnik and Alm (2002), and Ockenfels (2003). Resnick, Zeckhauser, Swanson, and Lockwood (2002) conduct a controlled field experiment using eBay's online reputation mechanism and survey the earlier literature. Brynjolfsson and Smith (2000) compared pricing behavior at 41 Internet and conventional retail outlets. They identify Internet sellers' trustworthiness as one important factor.

difficult to control for the many extraneous factors that come into play. For example, present technology is unable to wholly prevent people from changing their online identity and thus escaping their online reputation. So is a finding on a system's effectiveness due to the reputation system *per se* or because technology has not yet figured out how to prevent people from changing their online identity? The answer says a lot about how to go about improving things. But because the amount of ID fraud is presently anybody's guess, it's hard to say.

In the rest of this chapter, we describe a strategic framework for studying electronic reputation systems, one that was mapped out by game theorists in years prior to the Internet, but is nevertheless very useful for thinking about reputation and e-commerce. Our experiments test whether these systems work as theory says they should. Our laboratory experiments permit us to control extraneous factors, making it easier to puzzle out cause and effect. (We can add back the extraneous factors, in a controlled way, studying them as well). The lab also permits us to create parallel markets to compare reciprocity in long-term relationships to reciprocity under reputation systems–difficult to do under field conditions. Of course there is no free lunch; lab control sacrifices immediacy to the natural environment. But the insights gained in the lab can help us understand what we see in field, and may lead to ideas for improvements that would be difficult to come about in any other way.[4]

2. One-way reputation mechanisms: When the suppliers' trustworthiness is an issue

At base, reputation systems are intended to facilitate cooperation in situations where it would otherwise pay people not to cooperate. In a market setting, cooperation typically involves avoiding so-called 'moral hazards' like the false representation of the quality of goods or late payment (or maybe not paying at all).

To illustrate how reputation systems work, consider a simple electronic market with a typical moral hazard. There are many buyers and sellers in this market. For convenience, we think of the transactions as taking place over a series of rounds. Every round, each buyer and each seller goes to the market (actually, to their computers) with a budget of 35 thalers (a fictional currency). This is the total amount available for purchases or expenses for that round. Each seller offers to sell a single

[4] For detailed discussion on the complimentary relationship between experimental and field studies see Ariely, Ockenfels, and Roth (2002), Bolton, Katok and Ockenfels (2002), and Roth (2002).

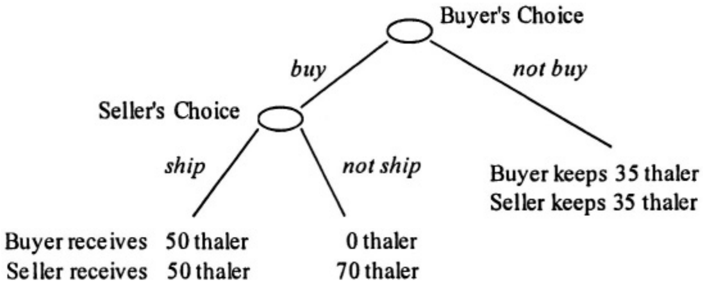


Figure 7.1. The buyer-seller transaction problem.

unit of some homogenous good. We will suppose this is a competitive market, so sellers are price takers; the good trades for 35 thalers. The good has a use value of 50 thalers to each buyer, and the seller's cost of providing a buyer with the item is 20 thalers. The decisions buyer and seller face each round are then illustrated in Figure 7.1. If the buyer chooses not to buy, both players keep their budgets. If he chooses to buy, he sends his 35 thalers to the seller, who then has to decide whether to ship the good, or whether to keep both money and good. If the seller does not ship, he receives the price plus his budget for a total of 70 thalers. If he ships, he receives the price minus the cost plus his budget for a total of 50 thalers, while the buyer receives his value of the good.

The moral hazard problem is shipping. If, as is common in electronic markets, the buyer-seller encounter is one-time or anonymous – buyer and seller are effectively strangers – the seller, once he receives the money from the buyer, has no monetary incentive to be trustworthy and to ship. Of course, some sellers may be principled and ship anyway. But those less attuned to principle, or those whose principle is their self-interest, may not ship. Anticipating this, and absent a way of distinguishing the trustworthy sellers, the buyer may decline to buy, so that trade does not take place, even though this would make everybody better off.

Of course, our electronic market is highly stylized (simplified). Still, it captures the essence of the moral hazard problem in several real-world electronic markets. Amazon.com, to give a prominent example, now offers used goods at its site right along side the new goods. Amazon actually acts as the middleman for the used transactions; the actual sellers range from brick-and-mortar stores to individuals selling personal items they no longer want. The site posting includes the purchase price the seller wishes to receive plus a description of the item and the shape it is in. A willing buyer sends the money to Amazon, which takes a cut, and passes the rest along to the seller. The seller is then supposed to ship the item within a pre-specified amount of time. While this market differs in several important respects from our stylized illustration (something we return to discuss in Section 4), the moral hazard problem is essentially the same (although in the Amazon case there are two seller moral hazards, the second having to do with describing the quality of the good).

Returning now to the stylized market, suppose rather than strangers, buyers and sellers are effectively partnered in a long-term relationship; that is, exclusively the same buyer has the opportunity to purchase from exclusively the same seller in every round. Intuitively, this changes everything; so too in formal theory. Partnered players are effectively playing a repeated game, permitting the buyer to condition his purchase decision on the past behavior of the seller. A simple discrimination strategy can provide the seller with an incentive to ship: The buyer buys if and only if the seller shipped in the last round. So long as the seller does not discount the value of future sales too strongly, the seller then has a monetary incentive to ship. So there should be trade. We are implicitly assuming that both buyer and seller are in the market endlessly, so that there is always future trading to consider. This is a simplifying assumption we will deal with in a moment. The sort of equilibrium in which players are cooperative now in order to elicit cooperation later is common to many repeated games and is known in the literature as a 'Folk theorem equilibrium' (to denote that no one knows who originally came up with the argument).

But electronic markets tend to involve trading among people and businesses that are more like strangers than they are partners – the reach of electronic markets is one of their main advantages. The question then is whether the kind of reputation strategy that can support trade in the partners market can also support trade in the strangers markets. In theory, the answer is that it can, and in these circumstances, rather easily at that. At first, this may seem surprising; we tend to think there is something special about direct reciprocity, about dealing with the same person over-and-over: I scratch your back so long as you scratch mine. But the message inside the Folk theorem is that the key to cooperation is not the players' *interaction* per se but rather the *information* that is available to the players. One quick way to see this is to suppose the partnered buyer has no short-term memory (much like the lead character in the recent movie *Memento*); so the moment a round of the market ends, the buyer forgets what happened. Obviously the incentive for the seller to be honest disappears with the information in the buyer's head.

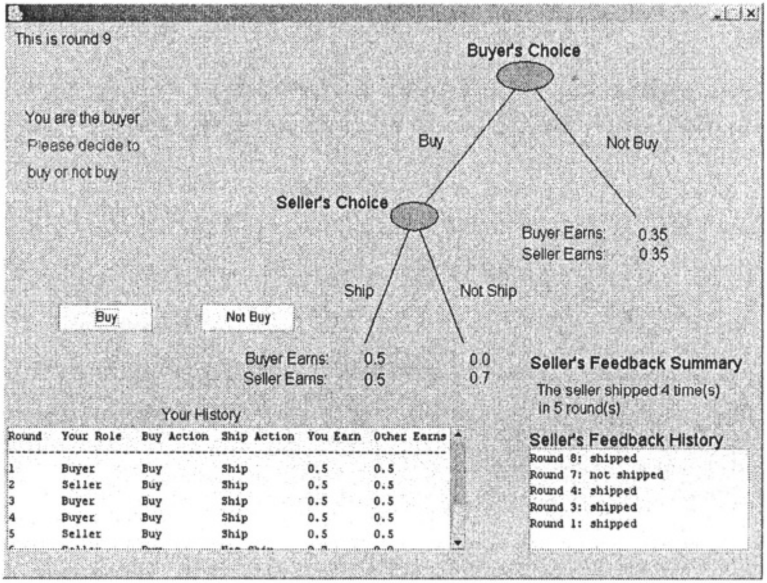
The information that the Folk theorem equilibrium effectively leverages is reputation. So long as a good reputation (in this case, a reputation for reliable shipping) is rewarded and a bad reputation is punished, the seller has an incentive to maintain a good reputation and avoid a bad one, independent of who's doing the rewarding or punishing. With incentives in place, the buyer can then be confident that a seller – even one unmoved by moral principle – will avoid falling into moral hazard. So in our stylized market, with strangers trading, providing buyers with a history of a seller's past shipping record should be enough to support trade.

In fact, this is precisely the insight that market platforms such as Amazon.com, eBay.com and Bizrate.com, among others, attempt to exploit. In theory, the system is a close, if not perfect, substitute, in the trust sense, for a one-on-one long-term buyer-seller relationship. (The field systems raise some additional issues we will ignore for the moment, concerning incentives to fill out buyer feedback forms and truthful reporting of feedback. We'll take them up in Section 4.)

But is the theory correct? That is, can a large group of strangers armed with information about reputation really trade as effectively as partnered traders? While plausible enough, it is a proposition that is virtually impossible to test in the field. For while there are sites like Amazon where strangers with reputation systems trade, there are few, if any, parallel markets involving just partners. We can, however, set up parallel electronic markets in the laboratory, and compare the performance of partners with that of strangers (with as well as without a reputation system). We did precisely this in Bolton, Katok and Ockenfels (2002).

In each of our experimental markets there were eight buyers and eight sellers. Each trader interacted with the market via a computer interface, meaning traders were effectively anonymous to one another (all were at computers in the same room, but these are separated by partitions so that no one can see what others are doing). The traders were Penn State University students, mostly undergraduates and studying in various fields. The market involved actual cash incentives: in fact, payoffs were as in Figure 7.1 save measured in U.S. cents.

The experiment consisted of three kinds of markets. In the *strangers markets,* buyers and sellers were randomly matched each round. No information about past histories (that is, about reputation) was made available. Traders in the *reputation markets* were matched in the same way but now, a buyer was shown feedback consisting of the shipping history of the seller he is matched with, prior to making a purchase decision. The computer interface of a buyer in this market is reproduced



This is round 9

You are the buyer
Please decide to
buy or not buy

Your History

Round	Your Role	Buy Action	Ship Action	You Earn	Other Earns
1	Buyer	Buy	Ship	0.5	0.5
2	Seller	Buy	Ship	0.5	0.5
3	Buyer	Buy	Ship	0.5	0.5
4	Buyer	Buy	Ship	0.5	0.5
5	Seller	Buy	Ship	0.5	0.5

Seller's Feedback Summary

The seller shipped 4 time(s)
in 5 round(s)

Seller's Feedback History

Round 8: shipped
Round 7: not shipped
Round 4: shipped
Round 3: shipped
Round 1: shipped

Figure 7.2. Buyer computer interface in the reputation markets.

in Figure 7.2. Traders in the *partners markets* were paired with the same partner in every round and so had the same feedback available to them as in the reputation treatment. Each kind of market was simulated three times. We never used the same traders in more than one simulation, so in total the experiment involved some 144 participants.

Each market simulation ran for 30 rounds, a fact that all traders were told up front. From theory, we would expect to see few trades in the final round or two – in all of the markets. To see this, consider again the game in Figure 7.1 and suppose that we are in the 30^{th} round of play. In terms of monetary rewards, the seller's optimal action, regardless of the game's history, is *not ship* since there are no future encounters there are no reputational benefits after round 30. A buyer who notices this should therefore be wary of buying. This remains true regardless of what seller feedback the buyer has, or of whether the buyer and seller are randomly matched or partnered. Good reputation has market value only if it can be leveraged in the future.

Note that if we push this reasoning to its ultimate end, *all* trade unravels: If the seller thinks there will be no trading in the 30^{th} round, there is no incentive to ship in the 29^{th} round, again independent of feedback or matching considerations, and so in the 29^{th} round the buyer should be wary of trade, ditto the 28th round, all the way back to the 1st round. There are several reasons, however, to think that things will not turn out this badly; some are theoretical reasons, others are behavioral. For one, it is unlikely that all buyers and sellers will see the unraveling argument we just gave. Some will just not reason it through while others will continue to trade out of principle. The thing is that even if you are able to reason it through, but some of your partners do not, then you may be better off *acting* as if you too do not. That is, if your partners think reputation is valuable for future rounds, then you are better off acting like it is too; at least for some time, at least until all players can see that reputation will not pay any more (Kreps, Milgrom, Roberts, Wilson, 1982). Studies suggest that many people have trouble looking ahead more than one or two steps, so from this we would expect less trading, but only in the last two rounds or so (e.g., Nagel, 1995, and Selten and Stöcker, 1986).

Of course, if as in the strangers markets, there is no information about reputation available, one need not reason far ahead at all to see the moral hazard problem: Since trade is always anonymous, one can see immediately that the seller has no monetary incentive to be trustworthy. So, in sum, theory leads us to believe that there will be more trading in the reputation and partners markets than in the strangers markets. Since, in theory, information is the only important thing to supporting trade, we expect to see about the same amount of trade in both reputation and partners, although in both cases we would not be surprised if the trade fell off in the last couple of rounds of the market.

The main results of the experiment, the trading rates in each kind of market, are displayed in Figure 7.3. First observe that trading rates in the strangers markets start fairly strong (about 45% on average), but quickly tail off to less than 10%. In these markets, but a few sellers are trustworthy, something buyers quickly pick up on. While trading in the reputation markets starts at about the same level as in the strangers markets, the trading levels remain high until the last two rounds, where, as suggested by theory, they crash. Overall trading levels are substantially higher in the reputation markets than in the strangers markets. But they are substantially higher still in the partners markets, hovering around 80 percent for most rounds until they too crash at the end. (All of the noted differences are significant at the 5% level, two-tail tests.)

Overall, the experiment indicates that there is some truth in the theory – but also some important problems. On the one hand, information about reputation does succeed in elevating trade. The strategic nature of reputation is evident from the low trading rates in both the strangers markets and in the last two rounds of all the markets. There are some

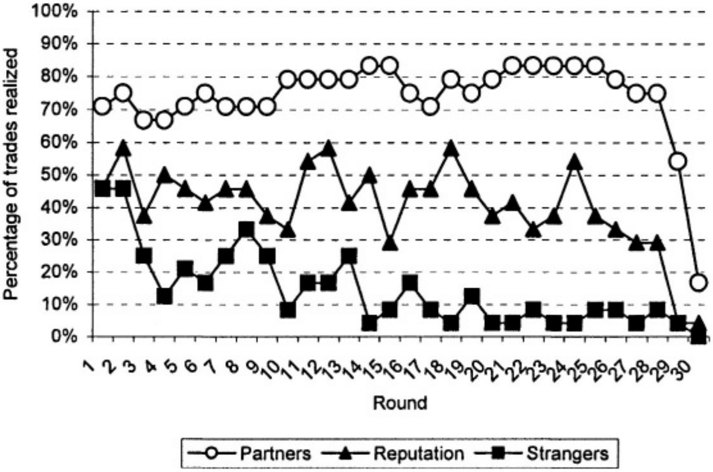


Figure 7.3. Trading in shipping markets, by round, averaged over all simulations.

people who are trustworthy on the basis of principle, but there are many who need an economic incentive. On the other hand, partnered traders manage to maintain substantially higher levels of trade. This might indicate there is something special about partnered relationships; we cannot entirely rule that out. However, further analysis (available in our paper) suggests that much of the difference is due to the differences in the way information flows through the two markets, providing the traders with different incentives. In both of these markets, a buyer choosing to trust can be thought of as investment; I trust to see if you are the kind of seller who is trustworthy. We have already discussed how a seller choosing to be trustworthy can be seen as an investment. It turns out that the return on these investments is higher in the partner treatments (trusting yields information about your partner; being trustworthy fills your partner with future trust) than in the reputation markets (trusting yields information useful to other buyers; some of the value of trustworthiness goes to a buyer you may never deal with again). These public good aspects of trust and trustworthiness in the reputation treatments - that are not captured by standard economic theory (see Bolton and Ockenfels, 2003, for a discussion) - prevent the electronic reputation system from being as effective as lasting relationships.

3. Two-way reputation systems: Markets where both sides' reputation is an issue

In many markets, reputation is a two-sided affair, with the trustworthiness of both buyer and seller at issue. It turns out that, in theory, these kinds of reputation systems, which track the reputation of both sides of the market, require a good deal more information to facilitate cooperation than do systems for one sided markets. Showing this in a full blown market setting, like the one we used in the last section, turns out to be rather complicated since other, superfluous issues get in the way. Specifically, in a two-way reputation system actions must always be taken in the appropriate context. For example, consider eBay's two-way reputation system where buyers rate sellers and sellers rate buyers. If a seller receives a negative rating from a buyer that says that the seller never shipped the product, this is not enough to condemn the seller because it could be that the seller did not ship because the buyer never sent the payment. However, even information about whether the buyer paid may still not be enough information because the buyer may not have sent the payment because after the auction completed he found out that the shipping costs were unreasonable. So in principle, the entire history of both trading partners as well as their trading partners, and their trading partners, and so on, may be required to construct a system that has sufficient information of the sort we tested in a one-way setting. Aside from being cumbersome, so much information is likely to be simply too much to process. It turns out, however, that the base issues can be captured by the extraordinarily simple 'image scoring game' (Nowak and Sigmund, 1998).

As simple as it is, the image scoring game is also ambitious. It aims to illustrate how reputation can be used to facilitate cooperation in just about any social situation, market or otherwise, where voluntary cooperation, and so the reputation of *all* participants is an issue.

As with the market we studied in the last section, the image scoring game conceives of the group interacting over a series of rounds. Again, in each round, people are paired off at random. Whether it is because the group is interacting anonymously (say, through computer interface) or because the group is large, we may assume that the people so partnered are strangers to one another. One person in the pair, designated the title of 'mover,' is given the opportunity to give a favor to the other, designated the title of 'receiver.' These designations are assigned randomly, so over many rounds, each player is a mover about half the time and a receiver the other half. Giving a favor would cost the mover c and benefit the receiver $b > c > 0$. Figure 7.4 illustrates the situation.

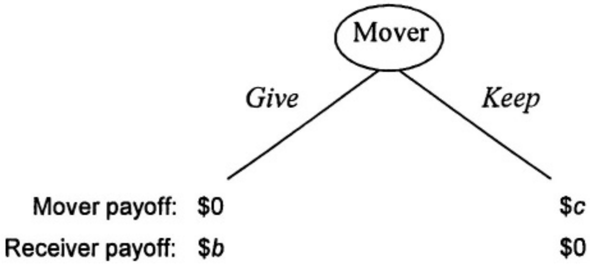


Figure 7.4. When Mover meets Receiver in the image scoring game.

But why would the mover want to help the receiver? The strictly self-interested answer is he wouldn't – unless, of course, giving now induced a future mover to give when he is the receiver. In fact, notice that the efficient outcome in this game, the outcome that maximizes the total social benefits, is for everyone to give when they are the mover. The rub is that if everyone else gives, then I make more money by keeping when in the role of the mover (while graciously accepting others' beneficence when in the role of receiver). This is where reputation can help, by providing the information necessary to reward those who give with giving and punishing those who do not with keeping.

At first, this game might seem miles away from the buyer-seller auction context we mentioned at the beginning of the section. But in terms of the reputation issue at stake, the two are actually quite similar. Both involve evaluating everybody's reputation, and cooperating, whether that means 'giving' or 'trading,' only if their reputation warrants it. The main difference is that in the market situation there is simultaneity to the evaluation – both buyer and seller evaluate the other's reputation at the same time. In the image scoring game, the evaluation is one-at-a-time. One-at-a-time makes for a more lucid discussion, but does not do violence to the basic reputation issue.

And the basic reputation issue is more complicated than before. To see why, consider the kind of discriminating strategy that worked so well for the one-sided case. It required a relatively small amount of information about the image score (reputation): the mover gives if he knows the receiver played give the last time as a mover, and keeps if the receiver last played keep. Before, this was a sound strategy for all, effectively curbing moral hazard and supporting cooperation. To see where it runs into trouble here, suppose you are the mover matched with someone who last player *keep* as a mover. Do you really want to play *keep* as the discriminating strategy stipulates? If you do, then the next time you are the receiver, you can expect the mover to play *keep* on you (if others too play the discriminating strategy). Consequently, you make more money playing give (lose c now, pays b later) than playing *keep* (gain c now, pays 0 later). The problem is that if enough people decide to give to keepers then it pays to be a keeper. And if it pays to have a bad reputation, then why have a good one? Why cooperate and give?

Of course, if enough people are willing to punish keepers, say out of a sense of social obligation, or perhaps because they do not think far enough ahead about the incentives, then cooperation may subsist on discriminating strategies in spite of the flaw. It would be nice if this were so because the amount of information about a person's reputation necessary for the system to function – what they did last time as a mover, something we will call *first order information* - is minimal. Another reason to suppose that it might work, is that the Maghribi trader system had a similar flaw – what incentive did merchants have to expend time, paper, ink and postage to expose a dishonest associate? – but the system nevertheless appears to have functioned well for many years.

But for the moment, suppose first order information is not enough. What information would be necessary to fix things? We could add *second order information* to the image score. Now the receiver's image score would include not only what he did last time as a mover, but also what the receiver he faced did last time as a mover. For example, the image score might state that the receiver "last played keep with a player who last played give." We call this second order information because of its recursive nature. This amount of information pushes the unraveling problem back by a step. To see this, consider a mover who, for the first time, encounters a receiver who played keep on a giver. To support his punishment, keeping on a keeper would have to be rewarded, meaning that there needs to be giving to someone who gives to a keeper – which is not consistent with self-interest since keeping on a keeper pays more. So now players would have to think *two steps* ahead, and be confident others do so as well, before cooperation would unravel. Is that enough? In fact, we can extend this line of thought to show that two steps is flawed if traders think three steps ahead. To prevent these problems you would need the *entire* transaction histories of the traders; not only the receiver's entire transaction history as a mover but also the complete history of all receiver's partners when they were movers. (The labeling mechanism that we discuss below is an attempt to create a processing mechanism that captures all this information and presents it in a transparent way.)

In theory, the only way to be really confident that cooperation will not unravel is if either a complete history of the game is available to all players, or if some sort of mechanism or institution is available to process and to provide the necessary information honestly (Milgrom, North and Weingast, 1990, Kandori, 1992, and Okuno-Fujiwara and Postlewaite, 1995). For this reason, some theorists have cautioned that indirect reciprocal systems might not be stable outside of very small groups where the information demands are relatively modest.

There are, however, reasons to believe that these systems are more stable than strict theory would suggest. As mentioned earlier, experimental research, much of it on prisoner's and other dilemma games, finds that people's ability to do backwards induction is rather limited, and that they tend to be myopic in their ability to look ahead. We might therefore conjecture that second order information is enough to support a discriminating strategy, or at least that second order information would yield more giving than first order information.

To find out, we conducted an experiment, running the image scoring game under three different information conditions: no information, first order information and second order information (Bolton, Katok and Ockenfels, 2003). Subjects were Penn State University students, mostly undergraduates from various fields of study, and recruited by fliers posted around campus. In total, there were 192 participants. We ran two image scoring games for each information condition, each game with 16 subjects playing for 14 rounds. Each round, subjects were anonymously paired, interfacing with one another via computers. The value of a gift, b, was $1.25 and the cost of giving, c, was $0.75. Subjects knew that they would be in each role, mover or receiver, for half the trials (7 times) and roles would generally rotate between rounds.

The main results of the experiment appear in Figure 7.5. When there is no information about reputation available, giving quickly tails off to rather low levels (19% averaged over all rounds). First order information greatly increases giving (35%) and second order information increases it further still (46%). The hypothesis that more information increases giving yields significance at the 5% level. Note in all cases, and as in the buyer-seller experiment of the previous section, giving tails off in the last two rounds pretty much independent of the information condition, again evidence of the strategic nature of reputation building.

But while information does improve cooperation, there is still considerable room for improvement. One thing we might try is to process *all* the information available for our players so that they might then apply a simple discrimination strategy in a way that cannot be cheated on (Kandori, 1992). For instance, in each round we might label each player as a

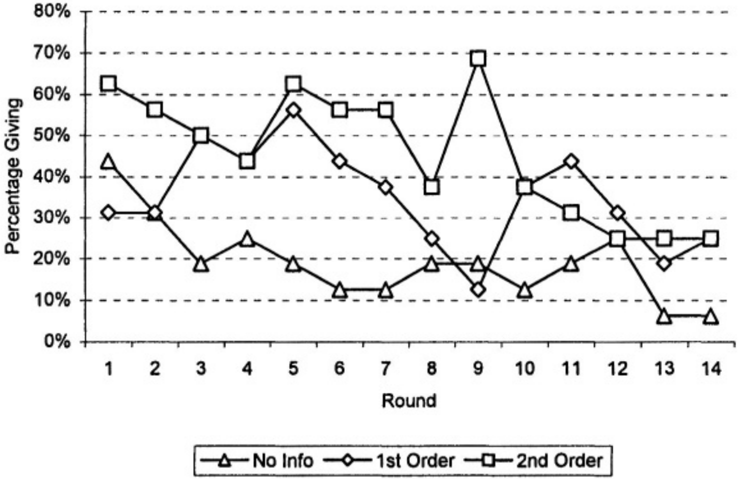


Figure 7.5. Cooperation in the image scoring game, by round, averaged over sessions.

member of either the ‘matcher’ club or the ‘non-matcher’ club according to the following rule:

- In the first round, everybody is a matcher.
- In every round after that, a player’s label is updated as follows:
  - If the player gave to a matcher the last time he was mover, he is a matcher
  - If the player kept on a non-matcher, he is a matcher
  - If the player did anything else, he is a non-matcher

Now consider a discrimination strategy that stipulates giving to a matcher and keeping on a non-matcher. If everyone follows this rule, then everyone will stay a matcher and there will be 100 percent giving. Moreover, you cannot benefit by cheating. If you keep on a matcher, you become a non-matcher which lines you up to be punished since the next time you are matched with a mover, he will keep on you. And punishment is now with impunity: keeping on a non-matcher allows a mover to maintain matching status – he won’t be punished for doing the right thing. A crucial property of this system is that giving to a non-matcher is punished in the same way as keeping on a matcher. This is necessary to insure that non-matchers are punished. When *all* information is processed in this way, the discriminating strategy looks pretty air tight. At least that’s the theory.

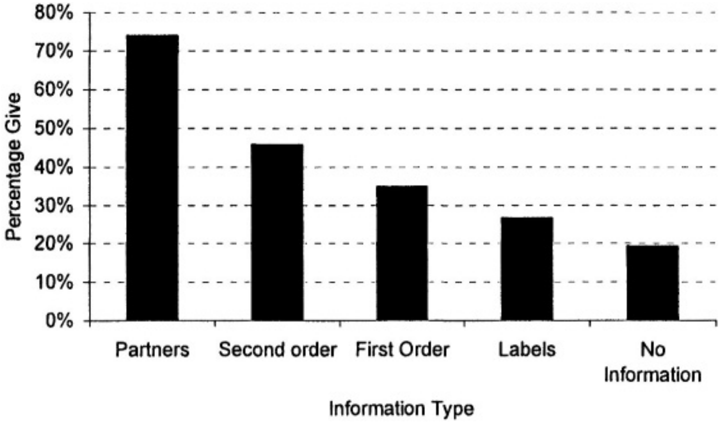


Figure 7.6. Giving levels in image scoring games under different information type.

It turns out, though, that this particular theory does not work so well in practice, or at least that is what further tests we ran seem to show. A summary of these tests is given in Figure 7.6, together with the previous results for the sake of comparison. We also ran a 'partners' image scoring game, quite similar in nature to the partners buyer-seller markets. As with the latter, partners induces far more cooperation than in any other version of the game. But the labeling scheme does but a little better than the no information games and somewhat less well than even first order information. That's a surprise, and one for which we have not yet been able to come up with a detailed explanation. There are other versions of the labeling scheme that are, in theory, just as effective as the one we described. We have tried a couple of these with no better luck than what you see in Figure 7.6. The most we can say is that it appears that people respond more favorably to reputational reports about recent past *actions* than they do to reputational reports that *filter* actions.[5]

Being matched with the same partner for the entire game induces more cooperation in the two-sided environment, just as it does in the one-sided environment. One reason for this is that in the partners treatment both players have the information for the entire game, so there is no ambiguity about the sequence of actions that led to the labels. Another reason is

[5] One such labeling scheme may be to display the entire history of the receiver's status as well as some summary information, such as the percentage of time he was a matcher (similar to the one-way system). We did not actually try to test this scheme because we think that it would not address the basic reason for the failure of the labeling system–people do not process filtered information as well as the information about actions.

that it is more transparent in the partners treatment that giving will lead to giving and keeping will lead to keeping in the future, since the reciprocity in this situation is direct.

4. Discussion and summary

In most supply chain transactions, successful buying and selling has traditionally been based on lasting relationships or secured by effective laws. The anonymous nature of online B2B environments and the opportunity of global and flexible exchange patterns in the Internet, however, tend to weaken the roles of repeated interaction and legal institutions. This poses an elementary problem to the effectiveness of e-supply chains unless mechanisms are developed that promote trust and trustworthiness and that function well even in adverse online environments.

Economic theory, field evidence and laboratory evidence all point in the same direction: The online trust gap can to some extent be filled with the help of online reputation mechanisms. Such mechanisms institutionalize word of mouth and thus make trust and trustworthiness profitable traits – even absent enforceable laws and personal relationships. Mutually beneficial trade among distant strangers in anonymous online markets is in principle feasible.

At the same time, however, our empirical research indicates that electronic reputation mechanisms do not easily substitute for partner relationships. The risk of trading with a cheater is in all our studies higher when reputation information is disseminated by an automated system than when it flows directly between permanent trading partners.

Our studies are all done in a highly controlled laboratory environment – but it's an environment that arguably favors reputation systems; that is, there is even more reason to worry about these systems outside the lab. These systems aim to curb moral hazard by collecting and disseminating the kind of information that is available to trading partners in personal, repeated relationships. But both collecting and distributing this information is more difficult in the field than it was for us in the lab. In particular, online reputation mechanisms typically rely on the *voluntary* provision of feedback information of experienced market participants. But when providing this information is costly, providing feedback is already a cooperative effort in itself, because the benefits of feedback information go to others. (Recall that because the informational benefit of a feedback goes to others, we also observe less trust in Bolton et al.'s (2002) reputation market than in the partner market.) In this sense, reputation mechanisms appear to shift the dilemma to another level rather than solving it – agents are supposed to cooperate with the reputation mechanism rather than with their trading partners. On eBay, for instance, sellers are only rated 50 percent of the time (Resnick and Zeckhauser, 2002). Furthermore, traders may have incentives to manipulate feedback information to, say, artificially raise a confederate's reputation, or to impugn a competitor. Finally, as noted in Section 3, the amount of information needed in a two-way system may be enormous and easily go beyond the scope of the traders' information processing capabilities.

We hasten to add that these problems are not the end of the story but rather the beginning. Ours and others' research offers promising ways to improve the effectiveness of automated reputation systems:

- The public good aspect of trust and trustworthiness identified in Section 3 can be weakened by inhibiting identity changes, and by informing all market participants about trustworthiness indicators of the whole market, and not only about the trustworthiness of individual traders.[6]

- Our findings point to the kind of information statistics that are sufficient, and those that are not needed to make the agents trust in each other. In particular, a cumulative measure of reputation, as applied by most online reputation services, does not seem to be appropriate because it hides information critical to the buyers' decision to trust.

- Incentives to provide information can be created through 'micro-payments.' If, in addition, these payments depend on the predictive value of the feedback for the future performance of the seller, then incentives to create honest feedback are obtainable (Miller, Resnick and Zeckhauser, 2002).

The research on automated reputation systems is just beginning, and the existing reputation systems on the Internet are still in their infancy. The emerging data suggest that there is plenty of room for improvement. But they also provide clues of how a clever architecture of electronic reputation mechanisms based on theory and empirical research might just be able to successfully bridge the trust gap.

[6] See Bolton et al. (2002) for the details, and see Friedman and Resnick (2001) for theoretical reasons why one should gain control over the agents' identities and how this could be technically realized.

Acknowledgements

Correspondence to Elena Katok, 509H BAB, Smeal College of Business, Penn State University, University Park, PA 16802; ekatok@psu.edu. Ockenfels gratefully acknowledges the support of the Deutsche Forschungsgemeinschaft through the Emmy Noether-program. Bolton and Katok gratefully acknowledge the hospitality of the Max Planck Institute for Research into Economic Systems during the summer of 2002. Katok gratefully acknowledges the support of the National Science Foundation award #DMI-0128588.

References

Alexander, R.D. (1987). The Biology of Moral Systems. New York: Aldine De Gruyter.

Ariely, D., A. Ockenfels, and A.E. Roth (2002). "An Experimental Analysis of Ending Rules in Internet Auctions." Working Paper, Harvard Business School.

Arrow, Kenneth (1974). "The Limits of Organization." New York: Norton, York.

Axelrod, Robert (1984). "The Evolution of Cooperation." New York: Basic Books.

Ba, S., and P.A. Pavlou (2002). "Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior." MIS Quarterly, 26, 243-268.

Bolton, G.E., A.E. Katok, and A. Ockenfels (2002). "How Effective are Online Reputation Mechanisms? An Experimental Investigation." Working Paper, Smeal College of Business Administration, Penn State University.

Bolton, G.E., A.E. Katok, and A. Ockenfels (2003). "Cooperation among Strangers with Limited Information about Reputation." Working Paper, Smeal College of Business Administration, Penn State University.

Bolton, G.E., and A. Ockenfels (2003). "Reputation, Information and Matching: Some Theory and Evidence." Work in progress.

Brynjolfsson, E., and M.D. Smith (2000). "Frictionless Commerce? A Comparison of Internet and Conventional Retailers." Management Science, 46, 563-585.

Dennehy, Michelle (2000), "Making B2B Trustworthy," special report posted on www.Auction.com.

Dewan, S., and V. Hsu (2001). "Trust in Electronic Markets: Price Discovery in Generalist versus Specialty Online Auctions." Working paper, University of Washington.

Federal Bureau of Investigation (2002). "New Report Shows what Internet Scams Cost Americans Most." Press Release on April 9, 2002.

Friedman, E., and P. Resnick (2001). "The Social Cost of Cheap Pseudonyms." Journal of Economics and Management Strategy, 10(2), 173-199.

Greif, A. (1989), "Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders," Journal of Economic History, 49, 857-882.

Houser, D., and J. Wooders (2001). "Reputation in Auctions: Theory and Evidence from eBay." Working Paper, University of Arizona.

Kandori, M. (1992). "Social Norms and Community Enforcement." Review of Economic Studies, 59, 63-80.

Kreps, David M.; Milgrom, Paul; Roberts, John and Wilson, Robert (1982), "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," Journal of Economic Theory, 27, 245-252.

Lucking-Reiley, D., D. Bryan, N. Prasad, and D. Reeves (2000). "Pennies from eBay: the Determinants of Price in Online Auctions." Working paper, Vanderbilt University.

Melnik, M.I., and J. Alm (2002). "Does a Seller's e-commerce Reputation Matter?" Journal of Industrial Economics, 50(3), 337-49.

Milgrom, P., D. North, and B. Weingast (1990). "The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs." Economics and Politics, 2, 1-23.

Miller, N., P. Resnick, and R. Zeckhauser (2002). "Eliciting Honest Feedback in Electronic Markets." Working paper, Harvard University.

Morphy, E (2001) "Putting the Strategy Back in Your Sourcing." iSource `http://www.isourceonline.com/article.asp?article_id=2049.`

Nagel, R. (1995). "Unraveling in the Guessing Game: An Experimental Study." American Economic Review, 85(5), 1313-1326.

Nowak, M.A., and K. Sigmund (1998). "Evolution of Indirect Reciprocity by Image Scoring." Nature, 393, 573-577.

Ockenfels, A. (2003), "Reputationsmechanismen auf Internet-Marktplattformen." Zeitschrift für Betriebswirtschaft, 73(3), 295-315.

Okuno-Fujiwara, M., and A. Postlewaite (1995). "Social Norms and Random Matching Games." Games and Economic Behavior, 9, 79-109.

Resnick, P., and R. Zeckhauser (2002). "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System." The Economics of the Internet and E-Commerce. Michael R. Baye, editor. Volume 11 of Advances in Applied Microeconomics. Amsterdam, Elsevier Science.

Resnick, P., R. Zeckhauser, J. Swanson, and K. Lockwood (2002). "The Value of Reputation on eBay: A Controlled Experiment." Working Paper, University of Michigan.

Roth, A.E. (2002). "The Economist as Engineer: Game Theory, Experimental Economics and Computation as Tools of Design Economics." Fisher Schultz lecture, Econometrica, 70, 1341-1378.

Selten, R., and R. Stöcker (1986). "End Behavior in Sequences of Finite Prisoner's Dilemma Supergames." Journal of Economic Behavior and Organization, 7, 47-70.

Chapter 8

STRATEGIES AND CHALLENGES OF INTERNET GROCERY RETAILING LOGISTICS

Tom Hays
United States Army
Huntsville, Alabama
tomhays90@hotmail.com

Pınar Keskinocak
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia
pinar@isye.gatech.edu

Virginia Malcome de López
Pegasus Group Inc.
San Juan, Puerto Rico
virginia@pegasuspr.net

Abstract One of the most challenging sectors of the retail market today is the grocery segment, specifically e-grocers. Since the mid-1990's multiple companies have entered the e-grocer market. Few have survived. What is it about e-grocers that make them fail? What makes them succeed? What can we learn from yesterday's e-grocers that will enable the new players to stay afloat or even be called "the greatest thing since sliced bread?" Given that the industry is still in "transition," it is difficult to find definite answers to these questions. With the goal of gaining useful insights on these issues, in this chapter we analyze e-grocers, past and present, discuss different business models employed, as well as order fulfillment and delivery strategies. Among the three alternative business models, namely, pure-play online, brick-and-mortar going online, and partnership between brick-and-mortar and pure-play online, we observe that the latter model has a higher success potential by combining the strengths and minimizing the weaknesses of the former two models. We discuss order fulfillment methods used by current and defunct e-grocers, specifically, the mega-warehouse, in-store order fulfillment, and the hybrid store-warehouse, and provide an overview of their relative advantages and disadvantages. We also discuss alternative order delivery methods, including, attended and unattended home delivery, in-store pickup, and third-party pickup locations. We conclude with a discussion on lessons learned from failure and success stories of e-grocers, a summary of current trends and directions, and future opportunities.

1. Introduction

Growing interest by consumers to point and click their way through nearly all aspects of daily life has fueled the Internet economy to develop services and sell products online even in areas that were once the sole domain of traditional businesses, such as grocery stores and pharmacies. Businesses that sell consumer products online have been coined as “e-tailers” and as “e-grocers” in the case of online grocery retailers.

Some of the reasons why an increasing number of consumers buy groceries online are common to all Internet purchases, including better prices, larger selection, convenience, and time savings. Home delivery of items purchased online is appealing to those for whom going out to shop is difficult for various reasons, such as physical disability, the need to care for small children, the lack of adequate or convenient transportation, and/or a busy lifestyle. Buying groceries and other products online unchains consumers from physically driving to and shopping in traditional stores. As George Shaheen, former CEO of Webvan, a now defunct e-grocer, stated, “we (e-grocers) promise to give back your Saturday mornings” (Anonymous, 2000b). This prospect is very appealing for many people, especially considering that the typical shopper visits a grocery store 2.2 times per week (Kahn and McAlister, 1997). In a study by McKinsey and Company, 82 percent of consumers who order groceries online do it as a substitute for frequent regular trips to a grocery store, rather than substituting for infrequent trips to stock up on limited items or for special occasions (Sneader et al., 2000). Hence, the industry could obtain the mass critical to success.

The value proposition of e-grocers is to become a single-source solution for the busy consumer today. E-grocers provide information as well as products and services. For example, on NetGrocer’s website, customers can preload a recipe onto the web page and with a simple click place an order for the ingredients of that recipe. Customization of weekly shopping lists and lists of frequently ordered products as well as personalized shopping aisles are additional benefits of shopping online versus in traditional stores. Storing weekly orders on the website makes repeat orders simple.

Despite its appeal not everyone has embraced online grocery shopping as a replacement to regular trips to the supermarket. In 2000, of the 4.5 million households that bought groceries online, only 1.1 million did so more than once a month. (Rubin, 2001). Having to plan and think about the delivery schedule are commonly cited reasons by customers as drawbacks for online grocery shopping. Most consumers are used to going to the grocery store on the spur-of-the-moment, and often combine grocery shopping with other activities including renting a DVD or video and dropping off or picking up dry cleaning. E-grocers' scheduling requirements and the desire to consolidate purchasing activities to a single trip are two factors affecting the consumers' desire to purchase groceries online.

Other factors negatively influencing consumers' decisions to purchase groceries or other goods online include shipping costs, credit card security, the need or want for immediate delivery of products, and the social aspects of the shopping experience. For most orders less than $50, e-grocers charge shipping costs ranging from $5 to $20 (see Table 8.1). Customers need to tradeoff the convenience of online shopping with the additional price paid on shipping. The threat of credit card theft remains a real issue in the minds of many consumers, although incidents of fraudulent credit card use on e-grocers' websites have been scarce to none. Some consumers are also concerned about the freshness of the products they buy from e-grocers, or they simply want to squeeze and smell the products before purchasing. The need to touch and feel products remains a barrier for the online sale of groceries, as well as other typically "high-feel" items like clothing or shoes. Although many customers do not think that ordering groceries online is something they would be interested in, once some try it, they become captivated by the convenience (Smaros et al., 2000). A more detailed analysis of consumer behavior and demand management can be found in Smaros et al. (2000).

Table 8.1. A summary of leading e-grocers, past and present.

Company	Items Sold	Delivery Fee	Delivery Schedule	Year Opened	Year Closed
Peapod.com	Perishable	$2.95 to $9.95 based on market & order size	Next day - 2wks	1989	
Netgrocer.com	Non-perish.	$4.99 to $599.99 based on zone & order size	FedEx usually 2-3 days	1995 1995	
GroceryWorks.com	Perishable	See Safeway.com		1999	Reconfiguring Jun-Nov 2001 See Safeway.com
Albertsons.com	Perishable	$9.95 delivery $4.95 pickup	90 min slots 10a-9p; same day deliv. after 5p for orders before 10a, next day pickup/ deliv. for orders before 12a	1999	
GrocerOnline.com	Perishable	$4.99-$599.99 based on zone & order size	UPS or FedEx 1-3 days based on zone	Founding company 1902	
Webvan.com	Perishable	$9.95 - orders < $75 $4.95 - orders $75-100 None - orders >$100		1999	July 2001
HomeGrocer.com	Perishable	None - orders > $75		1998	July 2001
Kozmo.com	Perishable/misc.	$2 for $5 min. order		1998	April 2001
Safeway.com Vons.com	Perishable	$9.95 - purchases ≤$150 $4.95 purchases >$150	2-hr slots, 8a-10p		2001
Shoplink.com	Perishable	$35 monthly		1996	Nov 2000
Streamline.com	Perishable	$30 monthly		1993	Nov 2000
FreshDirect.com	Perishable	$3.95 (1st time shoppers get $50 free & delivery fee for 1st 3 orders waived)	2-hr slots, 4-11:30p weekdays 9a-9p weekends	2002	
PublixDirect.com	Perish., DVD/VHS	$7.95 with $50 min. order	90 min slots 8a-9p daily	2000	
Potashbros.com	Perishable	$7-10	30 min slots, 3-8p daily, same day delivery - orders by 10a		

E-grocer sales have been growing steadily but still only account for a small portion on total grocery sales. None of the chains expects it to account for more than 5 percent of total sales ($400 billion in the US) over the long haul (Heim and Sinha, 2001). Yet online grocery sales are projected to be 46 percent higher this year (2003) from the previous, to $3.5 billion (Heim and Sinha, 2001). Further, the online grocery market in the United States is expected to grow from $600 million in 2001 to $4.9 billion in 2006 (Lee, 2003).

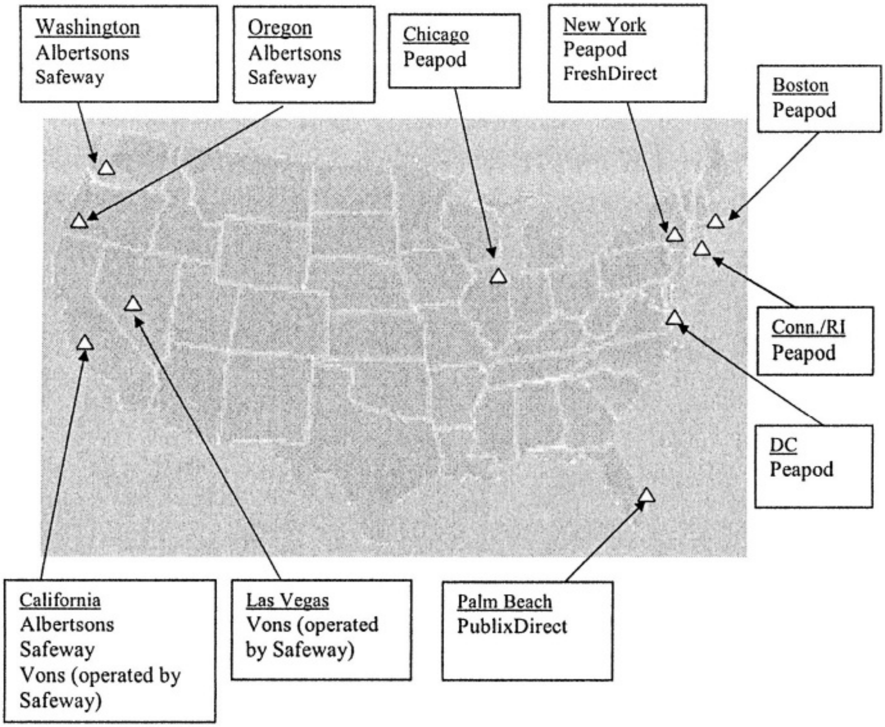
From a business perspective, the pure-play (no retail storefronts, only web ordering and delivery, and possibly one or more warehouses) e-grocer model has several advantages over the traditional retail grocery model. First, e-grocers do not have the high costs associated with multiple retail locations, including rent, parking, and high property taxes. Second, by "pooling" their inventory in fewer locations, e-grocers can better manage their inventory, reducing inventory holding costs and increasing inventory turnover rates. This leads to less spoilage of products and shorter lead times from the producer to the consumer, key advantages in selling perishable products. Third, e-grocers can collect detailed information about their customers' buying habits and preferences, which can then be used for targeted marketing and personalized promotions. Such one-to-one marketing is what the traditional grocers are trying to do by using loyalty cards to track purchases. Fourth, e-grocers may generate incremental sales growth for the industry. Although many retailers and consumers believe that impulse purchases diminish online, some companies such as Amazon.com have been able to foster impulse selling quite well with innovative marketing. Another sign that e-grocers can generate growth is an experiment the online grocer Streamline did with Procter & Gamble (Callahan, 2000). Streamline asked its customers if it could automatically add a toothbrush to the shopping cart every three months as the American Dental Association recommends. Customers liked not having to remember to replace their toothbrush. The computer kept track of it for them and created a large incremental increase in toothbrush sales.

Despite its potential advantages, the e-grocer model has its drawbacks. Brick-and-mortar stores have established locations, brand names, and a large customer base. A majority of consumers still prefer to buy groceries from a retail store. They may like to smell the vegetables and squeeze the fruit, or they may like to unwind from a long day by walking among the fresh breads of the bakery. Consumers also trust the grocery stores they have known for many years, and like to shop where they feel assured the quality is consistent and the price is right. But probably the biggest challenges for e-grocers are in order fulfillment and home delivery. Supermarkets aggregate demand by allowing customers to come to the stores, and therefore customers do the order picking and delivery. According to Procter & Gamble, traditional in-store shoppers who pick, pack, and deliver their own products now save the industry approximately 13 percent of the total cost of sales (Tapscott and Ticoll, 2000). For e-grocers, the more deliveries in a given area, the lower the costs per delivery. Except for a few cities, Americans who can afford computers and Internet access are more likely to live in suburbs, which means more driving and fewer deliveries per hour for any company that offers home delivery. In this chapter we present various models used by e-grocers, discuss the strengths and weaknesses in each, and provide an outlook for the future of the e-grocers. In Section 2 we discuss alternative e-grocer models. Sections 3 and 4 address order fulfillment and the methods by which the goods reach the customer. We offer our conclusions in Section 5.

2. Alternative E-Grocer Models

Major players in the e-grocery landscape differentiate themselves by the types of products and services they offer, particularly, by their method of order fulfillment and delivery (which is closely related to the types of goods sold), and by the geographical markets in which they operate (see Table 8.1 and Figure 8.1). All of these decisions are highly dependent on the business model of an e-grocer, i.e., whether it is (i) pure-play online, (ii) a brick-and-mortar company selling online, or (iii) a partnership/merger between a pure-play online and a brick-and-mortar company. In this section, we categorize and provide an overview of the major players in the e-grocer arena along these three dimensions.

The term "perishable" is used for those goods that are not pantry items and need to be stored in a refrigerator or a freezer. This category includes products like fresh vegetables and meats. Non-perishable products include all other household goods, including canned soup, canned fruits and vegetables, drinks, laundry detergents, and candy. Almost all e-grocers offer both perishable and non-perishable products (see Table 8.1). The goal of these companies is to replace consumers' trips to the grocery store completely. The main advantage of limiting the product selection to non-perishables is the ease of storage and delivery. But the trade-off is that customers would still need to go to a store to buy perishables such as milk, bread, and meat. To avoid having to shop both online and offline, some customers might prefer to buy only from those e-grocers who cater to all their grocery needs, while others may simply choose to go to the stores and avoid online purchasing altogether.



NetGrocer delivers non-perishables to 49 states (not incl. Hawaii).
GrocerOnline delivers non-perishable and frozen items to 48 states.

Figure 8.1. E-Grocer locations.

Currently, the main competitors in the e-grocer arena, selling both perishable and non-perishable goods, are Peapod, Albertsons, PublixDirect, and Safeway/GroceryWorks. NetGrocer delivers non-perishable, packaged groceries anywhere in the continental US using FedEx as its carrier.

A number of e-grocers, such as EthnicGrocer.com and LatinGrocer.com (also known as mexGrocer), have found and established themselves in niche markets such as specialty or ethnic food. Besides selling groceries, some of these companies also serve as a portal for the ethnic community they target, by posting events, news, community activities, etc.

2.1 Pure-Play Online E-Grocers

Several pure-play online e-grocers have entered, and left, the market in the past decade. In fact, the defunct e-grocers, including Webvan, HomeGrocer, Shoplink, and Kozmo, far outnumber the survivors. Most of the e-grocers who began as pure-plays have survived by merging or forming partnerships with other companies. For example, Streamline was sold to Peapod, another pure-play online, which was later sold to Ahold, a brick-and-mortar grocery giant. Players which fall into this category are discussed in Section 2.3. In this section, we focus on one of the newest and most innovative companies to enter the online grocer market, FreshDirect, and the most famous pure-play failure, the late Webvan.

Newcomer FreshDirect launched its grocery delivery service in New York in September of 2002. It expects $100 million in revenue for 2003 and projects $225 million for 2004 according the FreshDirects CEO Joseph Fedele (Kirkpatrick, 2002). With an initial investment of more than $100 million and 200 employees, FreshDirect hopes to turn an operating profit in six months (Dillon, 2002).

FreshDirect has a different product mix and value proposition than most online and brick-and-mortar grocers. Roughly, 83 percent of the foods sold by FreshDirect are perishables, compared to 20 percent in most grocery and 50 percent in the case of online grocer Peapod (Fabricant, 2002). Putting a very high emphasis on the freshness and the quality of its products, FreshDirect buys directly from the suppliers, farms, fishermen, etc. Storage, order processing, and food preparation takes place in its 300,000 square foot, twelve temperature zone warehouse in Long Island City, New York.

Defunct Webvan Group, Inc. was one of the leading e-grocers in 1999 and 2000, but closed its doors in July 2001 due to its failure to reach profitability. Webvan was headquartered in Foster City, California, was founded in 1998 by Louis Borders, who also founded Borders Books. Webvan introduced its service in the San Francisco Bay Area in June 1999 (Bellantoni, 2000). The company completed a $375 million initial public offering on November 5, 1999. In the first quarter of 2000, Webvan had 87,000 active customer accounts, up 85 percent from December 1999. The company had total sales of $259.7 million in 2000 (King, 2001).

In June 2000, Webvan agreed to buy HomeGrocer.com in an all-stock deal valued at about $1.2 billion. Before the merger, the companies had to compete not only with brick-and-mortar alternatives to win customers, but also with each other. The merger was designed to help the companies reduce marketing and customer acquisition costs and leverage their combined buying power with suppliers. The purchase reduced Webvan's capital needs by 50 percent, the company said, since HomeGrocer was already in several key markets, including Seattle and Los Angeles. Webvan instantly vaulted from two markets (three, including Sacramento as separate from San Francisco) to nine, and was expected to be in 13 markets by the end of 2000 (including Atlanta, Baltimore, Bergen County, New Jersey, Chicago, Dallas, Los Angeles, Orange County, California, Portland, Oregon, Sacramento, San Diego, San Francisco, Seattle, and Washington, D.C.) and in 15 markets by mid-2001.

Investors did not see the union positively, and the stock prices of both companies dropped sharply after the deal was announced. While long-term prospects for the merger had good potential, investors felt that in the near term the new company faced a host of hurdles. The two companies (i) had overlapping market expansion plans that had to be rectified, (ii) used different technology platforms, and (iii) had different approaches to logistics. Webvan preferred large distribution centers and a 30-minute delivery time, while HomeGrocer.com used smaller centers and delivered within 90 minutes. HomeGrocer's smaller distribution centers cost about $5 million each to build. They were far less automated than Webvan's $35 million warehouses, and in fact, the company tried to build a warm and fuzzy image around the idea of human picking and packing.

While the two companies struggled to decide on which business model would survive, Webvan's web pages replaced HomeGrocer's web pages. This change, apparently thought to be cosmetic and largely transparent to the customer, caused a one-third drop in demand from HomeGrocer's. The switching cost of learning a new web page and the difference in delivery policy, plus technical problems with the web page was more than some customers wanted to bear.

At the end of 2000, Webvan averaged 2,160 orders a day in the San Francisco Bay area. The break-even point was close to 2,900 orders per day, which meant that Webvan was operating at more than 25 percent below its breakeven point. In March 2001, the average customer order was $114. In April 2001, Webvan said that it was breaking even in Orange County. However, this was not enough to gain investors' confidence. Between February and June of 2001, Webvan's stock plummeted. Louis Borders and George Shaheen left Webvan, operations were closed in three cities (Atlanta, Sacramento, and Dallas/Ft.Worth) and 1,150 employees were laid off. Webvan never recovered and closed its operations in July 2001 (Knowledge at Wharton, 2001). Webvan and Homegrocer combined raised a total of $1.2 billion in capital in their short lives and used it all (Lee, 2003; Kane, 2002).

Comparison of the business models employed by Webvan and FreshDirect demonstrates sharp contrasts. While FreshDirect chose to stay focused, Webvan followed a quick expansion approach both in terms of the markets served and the products and services offered. It is now evident that the latter approach had several problems. It required large investments in multiple markets, before the concept and implementation was fully tested and refined in any market.

2.2 Brick-and-Mortar Grocers Selling Online

Large brick-and-mortar companies have been slow in entering the e-grocery market, but fared better than pure-plays. Traditional grocery stores did not initially perceive the need to offer online ordering and delivery service, but the entrance and perceived success of pure-plays caused them to reevaluate their strategy. The benefit for the traditional stores is the retaining of their original customers and the possibility of tapping into a new market. The clicks-and-bricks strategy gives customers the flexibility of shopping the brick-and-mortar store, as well as ordering via the Internet and to pick up their order at the store or having the order delivered. Customers can also place their orders using store-based computers. The most prominent brand-name grocery stores now online are Albertsons, Publix, and Safeway. Safeway's online store is a partnership with the former pure-play firm Groceryworks.com, and also Tesco, therefore we discuss both Tesco and its partnership with GroceryWorks/Safeway in Section 2.3. Also briefly mentioned are some e-grocers in the Asian market.

Headquartered in Boise, Idaho, Albertsons was founded in 1939. Currently, Albertsons is the second largest supermarket company in the United States, and its online store Albertsons.com serves more than 700 zip codes in California, Oregon, and Washington giving it the largest geographic reach of any online grocery provider (Albertsons, Inc., 2002). Albertsons.com was launched in November 1999 in Seattle, expanded to San Diego in October 2001, and then to Southern California in March 2002. Albertsons has an annual revenue of $36 billion, and employs 200,000 associates in 2,300 stores. Although the percentage of revenue from the online sector was not available, the first quarter earnings this year (2003) are up from last year's loss by $337 million. All goods offered in stores are available on the website at the guaranteed same price. Online orders are fulfilled by Albertsons.com via existing stores. Pickers, dubbed "e-shoppers" by the company, fill orders by shopping the aisles of Albertson stores alongside regular customers.

PublixDirect, headquartered in Alpharetta, Georgia is a wholly owned subsidiary of Publix Super Markets Inc. Publix operates 759 stores in Alabama, Georgia, Florida, Tennessee, and South Carolina, with 2002 retail sales of $15.9 billion. Publix is one of the top ten largest volume supermarket chains in the US, with 119,500 employees. In October 2001, PublixDirect began accepting online orders for grocery delivery in several areas of Palm Beach and Broward counties in Florida. They have now expanded to include 106 zip codes in Broward, South Palm Beach, North Dade, and Key Biscayne. They also offer non-grocery products such as DVD, VHS, and flowers.

In Asia, clicks-and-bricks grocers include Fairprice and Cold Storage in Singapore, and Wellcome Supermarkets and Park'n Shop in Hong Kong. FairPrice has a network of more than 90 stores island-wide. Owned by more than 400,000 Singaporeans, it had sales of $1 billion in 2000 and employs a staff of 4,000. FairPrice was the first supermarket retailer in Singapore to own its own central warehousing and distribution system and named it the Grocery Logistics of Singapore. Cold Storage operates 35 supermarkets in Singapore and is a wholly owned subsidiary of Dairy Farms, a large supermarket, convenience store, and drugstore conglomerate. In 1997, Cold Storage began its Dial and Deliver service, allowing delivery of customer orders placed via the Internet, telephone, and fax within 24 hours. The chain has an aggressive expansion plan and delivers to all stores through its Fresh Food Distribution Center.

In Hong Kong, Wellcome is a wholly owned subsidiary of Dairy Farms (see above), which operated 239 stores as of June 2001. Wellcome delivers groceries, including frozen foods, alcohol, and tobacco, to Hong Kong Island, Kowloon, and New Territories. Park'n Shop is a subsidiary of A.S Watson Group (HK), a large retail and manufacturing company with a strong presence in Mainland China, Hong Kong, Macau, Taiwan, Singapore, Malaysia, and Thailand. Park'n Shop has over 190 stores and

9,000 employees in Hong Kong and delivers groceries, including frozen foods, baked goods, and miscellaneous general merchandise (over 4,000 items).

Hong Kong and Singapore seem to lead the western world in the e-grocer arena by quickly assimilating technology into everyday life. Both of these countries are very prominent in terms of Internet connectivity and have a high density in population. The former characteristic makes it easier for customers to adopt the e-grocery concept, while the latter provides both advantages and challenges for grocery delivery. In both of these countries, owning a car is not nearly as common as it is in the United States, hence customers often need to use public transportation for shopping. Having groceries delivered to their door rather than carrying them in crowded buses or subways is a big advantage. From the e-grocer's perspective, it is advantageous to have a large number of deliveries in a relatively small area, but making timely deliveries might be a challenge due to heavy traffic. In these respects, these two Asian countries share similar characteristics with New York City, where FreshDirect has been operating successfully.

2.3 Partnerships Between Pure-Play Online and Brick-and-Mortar Companies

Partnerships or mergers between pure-play online and brick-and-mortar companies have been increasing recently. Brick-and-mortar supermarkets have a large, loyal customer base and an efficient logistics system that allows them to make profits in the narrow-margin grocery business. Along with their infrastructure, brick-and-mortar companies often have tremendous brand name recognition and financial backing, allowing them to try different channels to reach new customers, and reducing marketing/advertising costs to attract customers online. Such costs can be prohibitively high for pure-play e-grocers; for example, Webvan spent between 25 and 35 percent of its revenue on advertising, compared with about one percent for traditional grocery chains, according to Rob Rubin, an analyst at Forrester Research (Moore, 2001). The high fixed costs of building a logistics infrastructure for order fulfillment and delivery discourage investors who are looking for quick profits. Partnerships between the brick-and-mortar companies and pure-plays create clicks-and-bricks companies that combine the strengths and minimize the weaknesses of both types of models.

2.3.1 Peapod with Royal Ahold. Founded and headquartered in Chicago (Skokie), Illinois, Peapod is one of the largest e-grocers and first to the market, delivering directly to customers' homes in major metropolitan areas including Boston, Chicago, Greater Washington D.C., southern Connecticut, and Long Island. Peapod expanded its markets and increased its sales fairly quickly, but at the same time its expenses also rose significantly. After unsuccessful attempts to raise financing in the public markets, Peapod sold a 51 percent stake in April 2000 to Royal Ahold, for $73 million, or $3.75 a share (Peapod, 2000a). Royal Ahold is an international leading food provider with annual sales that reached nearly 63 billion euros in 2002, with U.S. retail sales of over $26 billion (Ahold, 2002b). In September 2000, Peapod-Ahold agreed to acquire the Washington D.C. and Chicago assets of Streamline.com for $12 million. Streamline's operations switched to the Peapod brand in the Baltimore-Washington area in the fourth quarter of 2000. Simultaneously, Peapod withdrew from the Texas and Ohio markets. Streamline's Chicago assets served to help Peapod expand in that region. In July 2001 Royal Ahold offered to buy the remaining 42 percent of Peapod for $35 million. After the deal closed in August 2001, Ahold owned over 90 percent of Peapod. The purchase of Peapod by Ahold has brought this e-grocer completely under the wing of an existing brick-and-mortar grocery giant. In July 2001, Peapod in Chicago reported an operating profit. But, order growth was not rapidly increasing and high costs associated with logistics made Peapod vulnerable to ceasing its operations because of negative cash flow. By the end of 2001, Peapod reported an annual operating loss of $47.9 million on turnover of $100 million. In its third quarter net earnings report for 2002, Ahold reported a reduction in losses by Peapod with only $7.4 million in 2002, compared to $11.1 in 2001 (Ahold, 2002a).

As in other pure-play online and clicks-and-bricks alliances, what Peapod and Streamline bring to the partnership is e-commerce and home shopping expertise, web-based software and ordering systems, web marketing and additional information technology (IT) skills. Ahold's contributions lie in its considerable buying power, real estate, strong store brand recognition, extensive customer base and category management expertise.

2.3.2 GroceryWorks.com with Safeway and Tesco. Founded in 1999, GroceryWorks is headquartered in Dallas, Texas, and operates in Dallas, Fort Worth, Houston, and Austin. In June 2000, Safeway invested $30 million for a 50 percent stake in GroceryWorks (Sandoval, 2002a). Safeway operates more than 1,700 stores with annual sales of $34 billion (Sandoval, 2002a). Under the terms of this arrangement, GroceryWorks became Safeway's Internet division, and the two companies brought online grocery shopping to markets where Safeway operates. In September 2000, GroceryWorks and Randalls Food Markets partnered to provide home delivery to Austin area residents. The service operated in Austin as "GroceryWorks.com by Randalls." On June 25, 2001, Tesco, the British food giant, acquired 35 percent of GroceryWorks voting stock and raised $35 million in additional financing (Anonymous, 2001; Sandoval, 2002a). Today, the strategic partnering of the two food titans, Safeway and Tesco, with GroceryWorks is making a significant impact on the e-grocer industry.

Going to www.groceryworks.com automatically brings the customer to the Safeway order and delivery website, www.safeway.com with service to Northern California, Oregon, and Washington (GroceryWorks, 2003). In addition to these two sites, Safeway also operates online grocery services via Vons.com to Las Vegas and San Diego. Vons.com also uses Groceryworks software and Tesco know-how like its sister site, Safeway.com.

Tesco was founded in 1924 and became the largest food retailer in the U.K. by 1995. Tesco Direct, the online operation of Tesco, claims to be the biggest online grocery business in the world. In 2000, Tesco Direct had nearly 300,000 registered customers and 2000 sales of more than £250 million (Hodges, 2000; Wallace, 2000). In 2001, it has "almost one million registered customers, annual sales of about 300 million pounds, and is profitable" (Muehlbauer, 2001). Tesco argues that it makes a respectable margin on its Internet sales, because the average web-shopping basket is worth about £100, much more than people normally spend when they shop in stores. To improve revenues, Tesco expanded its product selection beyond groceries to include higher margin products, such as books, CDs, DVDs, games, flowers, baby items, home furnishings, and clothes that online customers seem more willing to mix with food than do customers in the supermarket. This is a similar strategy that Webvan attempted to follow. "Amazon developed a book business and then diversified into a whole load of other services, including groceries," says Tim Mason, Tesco's e-commerce chief (Hodges, 2000). "We're doing it the other way around."

As part of its web strategy, Tesco also offers a number of non-grocery services. For example, Tesco.net, the supermarket's free Internet service provider, has around one million members and ranks among Britain's largest ISPs. Once a customer is connected to Tesco's portal, she could buy one of 1.2 million books from Tesco's branded online store or choose from a stock of thousands of CDs and videos. Alternatively, she could click on Tesco's financial services Website, a joint venture with the Royal

Bank of Scotland, and apply for a Tesco home-insurance policy, a Tesco savings account, or a Tesco Visa credit card.

2.3.3 NetGrocer.com with Various Supermarket Chains. Founded in 1996, privately held NetGrocer.com offers 7,000 SKUs and enables shoppers nationwide (48 contiguous states) to purchase nonperishable groceries, drug store items, and general merchandise online and have them delivered to their homes and offices. Netgrocer.com's largest shareholders are Parmalat S.p.A. and Cendant Corporation. Parmalat S.p.A. is a subsidiary of Parmalat Finanziara S.p.A., a listed company incorporated in Italy. Its business is concentrated in milk and its derivatives, fruit juices and bakery products. Cendant Corporation is a global provider of financial and customer services.

NetGrocer.com is now a division of NeXpansion, Inc. and expanded its offerings to include e-commerce solutions for manufacturers and retailers, in addition to its primary business as an online grocer. As explained on their website, nexpansion.com, NeXpansion has partnered with numerous supermarket chains, including Lowes Foods, Harris Teeter, Pathmark, Big Y, Clemens Family Market, and Stop & Shop to offer customers hard-to-find and specialty items. NeXpansion provides a customizable web solution to these firms called "Endless Aisle." This deliverable allows customers to access the web via home computers or in-store kiosks, log on to the store's website, and from there search for items not available in the store. The items are then shipped to the customer.

3. Order Fulfillment

The trend for order fulfillment among most pure-play online e-grocers has been to establish large, automated distribution centers for each major market they serve, while some brick-and-mortar chains have employed only in-store order fulfillment. In addition to these two alternatives, consolidation in the market and the entry of existing brick-and-mortar stores into the e-grocer segment lead to a hybrid model that combines fulfillment from existing stores and smaller fulfillment centers. We discuss each of these models in this section (see Table 8.2).

3.1 The Mega-Warehouse

The mega warehouse model is employed by several e-grocers. NetGrocer began with a large, automated warehouse, a football-field sized structure to house all its non-perishable groceries, electronics, books, and music. Webvan also began with a huge, $20 million warehouse, to host its Oakland operation (Richtel, 1999), as did HomeGrocer and ShopLink. FreshDirect has a 300,000 square feet facility which hosts 12 different temperature zones to keep produce fresh and avoid contamination. Due to FreshDirect's focus on food preparation, the facility in Long Island City, New York appears to be more of a food processing center, or giant kitchen, than a warehouse or distribution center. Park'n Shop of Hong Kong has six regional computerized distribution centers. The company committed considerable capital investment to advanced food technology and owns Asia's first multi-temperature distribution and processing centers, built at a cost of $30 million. This state-of-the art facility provides an unbroken cold chain for fresh, chilled, and frozen foods from source to customer.

For most of the e-grocers, very limited information is publicly available on warehousing and order fulfillment. We had the opportunity to visit the Webvan distribution center (DC) in Suwanee, Georgia, and also talked to several of their logistics managers; hence, in the remainder of this section we focus on Webvan's distribution centers and operations. For an in-depth discussion on designing, operating, and managing warehouse facilities, we refer the reader to Bartholdi, III and Hackman (2003).

Webvan made a strategic decision to build massive, highly automated warehouses with sophisticated inventory software, for $35 million each. In July 1999, Webvan contracted with the construction firm Bechtel to build up to 26 of these centers for $1 billion. Among the most publicized was Webvan's 336,000 square-foot distribution center in Oakland, California, 42 times the size of the fast-pick in-store fulfillment centers being rolled out by Ahold. Webvan's fully automated and temperature-controlled DCs allowed for processing 50,000 SKUs and roughly the volume of 18 supermarkets. Orders were processed with proprietary software, using automated carousels and conveyors for order picking. DCs were filled with miles of conveyor belts carrying bins. These bins, as well as the different zones in the warehouse, were color-coded: yellow for ambient products (which do not need refrigeration), green for chilled goods, and blue for frozen products. There were also deli and produce areas and a butcher, where meat was cut to order.

To fulfill an order placed by a Webvan customer online, a computer launched a set of multi-colored bins that correspond to like colored zones in the warehouse. Employees stood at one end of fifteen-foot high rotating racks, picked items off the racks and placed them in their respective bins on the conveyor belt (see Figure 8.2). Employees knew which item to pick via a computerized system of lights. The system illuminated by electronic display what rack the proper items were on, and which items should have been placed in which bin. A network of computers and scanners were controlled by logistics software from Descartes and Harbinger, which coordinated the movements of the racks of products and the conveyor belts. With these warehouse management and automated pick and pack systems, Webvan claimed that the 150 or so workers in the Oakland facility never had to move more than 19 feet to fill an item in an order.

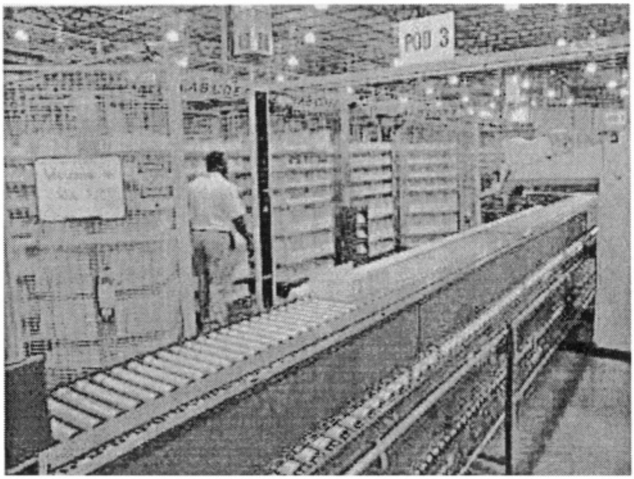


Figure 8.2. Carousels at Webvan's automated distribution center in Suwanee, Georgia.

To reduce the delivery costs to customers and increase delivery timeliness Webvan delivered its products to consumers via a "hub-and-spoke" distribution network. In the DC, the color-coded totes holding customers' orders were grouped and sorted by order number and destination, and loaded onto carts, which were then loaded onto trucks and shipped to one of the "stations" near the final delivery locations. The carts (loaded with totes) were cross-docked from trucks to vans in the stations and the vans delivered the orders to customers (see Figure 8.3). With this mechanism, each tote was lifted manually only twice after being filled: when it was loaded onto the cart in the DC and when it was carried from the van to a customer's home. This network centralized the order fulfillment and decentralized the delivery system, providing a more cost- and time-efficient process in conquering the last mile of e-commerce.

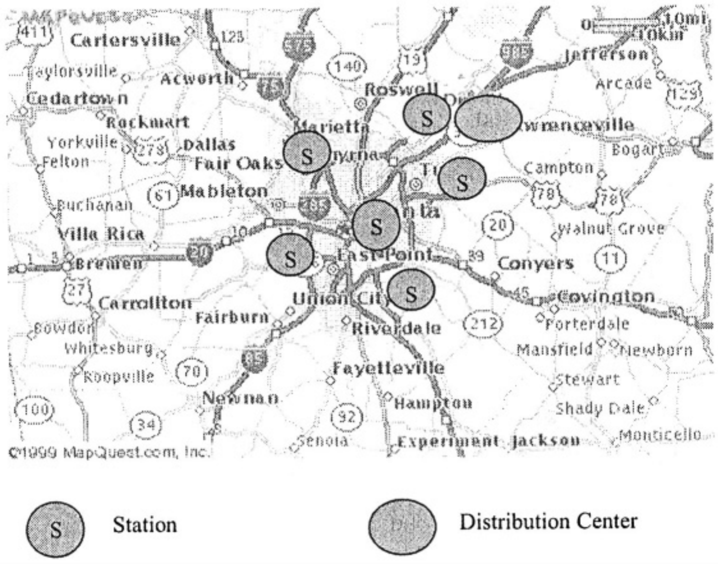


Figure 8.3. Webvan's hub and spoke delivery network in Atlanta area.

Webvan entered multiple markets simultaneously to gain the coveted first mover advantage. But such a large initial investment in so many markets, without a proven business model, resulted in significant financial challenges. The high fixed cost of implementing the warehouse and inventory management software coupled with high facility construction costs gave Webvan a high breakeven point for its sales. Unfortunately, the expectations and forecasts of demand were overly optimistic, and Webvan's facilities were only operating at half capacity, making it impossible to reach breakeven.

There are several advantages of a large, centralized distribution center. Inventory is centralized, leading to higher turnover, lower inventory costs, and fresher products. Delivery costs from the suppliers are lower, since deliveries are made to a single location in higher volumes. Order picking and packing can be automated using high-tech warehousing equipment and systems. Unlike the "pick and deliver from stores" model, a large automated DC leads to lower labor costs, is more efficient, and scalable for larger volumes . The main disadvantage of the mega warehouse model is that the distribution centers are very expensive to build, costing $25 million to $35 million each. To realize cost savings, capacity utilization must be high and not vary significantly over time (Kamarainen, 2003). The cost of experimenting on such a facility is very high, and in case a mistake is made in the design, it can have disastrous effects. Cost of delivering to homes also can be prohibitive, since such big warehouses usually need to be built in distant locations. To mitigate this drawback, some e-grocers used a hub-and-spoke system for deliveries, as discussed above. In short, most e-grocers still do not have the critical mass of customers that is needed to attain profitability with this model, and hence, a flexible distribution center which relies on manual solutions rather than automation could be more viable.

3.2 In-Store Order Fulfillment

The trend among brick-and-mortar companies entering the online grocer market is to use their existing facilities to fulfill orders. Our discussions of Tesco and Safeway (GroceryWorks) in this section describe how just a few of the growing number of clicks-and-bricks grocers leverage their stores for order fulfillment.

Tesco's fulfillment and delivery model resembles the early days of Peapod. Instead of building highly automated new warehouses dedicated to filling online orders, the strategy of rival Sainsbury's, Tesco is exploiting its network of nearly 700 stores nationwide (in the U.K.). Tesco employs teams of people to pick the items on their customers' web-transmitted shopping lists from the shelves of the nearest supermarket, and teams of drivers deliver the orders at agreed times. When an order is received from the Tesco Direct website, it is routed to the nearest physical outlet. For assembling an order, employees use special carts (called "picking trolleys") mounted with screen guides and "shelf identifier" software instructing them where to pick the items in the list (Sandoval, 2002a; Beck, 2000). Once the trolley is loaded, it goes straight to the van for delivery (Maddali, 2003; Millenium, 2003).

With its current fulfillment and delivery model, Tesco has been able to develop its online business faster than its competitors who are creating parallel distribution systems from scratch. But there are doubts about the profitability and the scalability of its web business. Tesco's fulfillment model is very labor-intensive. The £5 Tesco charges for home delivery may not be enough to cover the cost of employing pickers and drivers. Recently, the cost of picking and home delivery operations at Tesco have been estimated at 14 percent of sales, divided evenly between both operations (Reinhard, 2001). Moreover, when stores are crowded at weekends, which are when most orders arrive, efficiency falls off further as pickers jostle their way down aisles and queue like any other shopper. Therefore, to avoid a potential competition between regular shoppers and pickers in busy stores, it might be more efficient to do the order picking from less-busy stores rather than from the nearest store to the consumer (Kumar, 2001). In assigning orders to stores an e-grocer needs to tradeoff the picking efficiency with delivery distances, times, and costs. Finding the "best" picking location for an order is an interesting research problem. It could potentially be modeled as a variant of the multiple-depot vehicle routing problem with time windows (Laporte, 1992; Desrosiers et al., 1995), where there is a "congestion cost" at each retail store (depot) which increases in the number of orders assigned to that store. The goal is then to assign the orders to the stores and construct delivery routes for the vans such that the total congestion and travel costs are minimized, subject to constraints such as the capacities of the vans, the delivery time windows, the number of pickers (or the picking capacity) available at each store, etc.

Through its short history, GroceryWorks was invested in and bought partially by Safeway, then Tesco. GroceryWorks' business model changed several times when the new investors (buyers) purchased their portions of the company. The model started with one large warehouse, changed to multiple, smaller warehouses, and it is now in-store picking. At the end of June 2001, GroceryWorks announced that it was closing all of its Texas distribution centers and now has a new business model of delivering completely out of existing stores under the Safeway and Tesco umbrellas (Kane, 2002). The emergence of this new model allows GroceryWorks to reduce costs by not building and staffing its warehouses.

3.3 Hybrid Store-Warehouse

The third model of order fulfillment seems to be a natural evolution from the previous two. Picking from existing stores is a good alternative if fast roll-out with low investments is required (Yrjölä, 2003). The in-store pick model is also appealing to brick-and-mortar supermarkets cautiously entering the market. Leveraging existing assets lets them test the market without spending a great deal on new facilities. On the down side, in-store picking is naturally more inefficient than having a warehouse with shelves and aisles strategically planned for quickest picking and packing time. Therefore, when the market has developed, the customer base has been established, and the proper density is achieved, a plausible alternative is to move to a distribution center and for a time, maintain both in-store picking, either straight from the aisles or from a dedicated non-customer area in the store, and a warehouse. The three companies discussed below, Sainbury's of U.K., Peapod, and Albertsons, are traversing this path.

Sainbury's evolution followed the path described above: orders were originally picked and packed in-store by a team of personal shoppers and customers could pick up their orders from the store at a service charge of £3.50 or have it delivered to their home at an agreed time for a charge of £5. In May 1999, Sainsbury's announced that it will open the U.K.'s largest food picking center to fulfill customer orders received online. Sainbury's has since built two picking centers and stores nearly 15,000 products in these picking centers. The decision was based on Sainsbury's belief that while a store-based system is operable in principle, it is neither viable nor capable of dealing with significant volume without affecting the quality of services being offered to shoppers in-store. Currently, Sainbury's employs a hybrid-picking model based on two dedicated picking centers and in-store-picking in 33 stores. By moving to this hybrid fulfillment center/in-store picking model, Sainsbury's forges new ground and differentiates itself from its rival Tesco (Davis, 2001).

Peapod began its service by having personal shoppers go to grocery stores and fill orders made online by picking items right off the shelves. Orders were then delivered by mini-vans to customers. Peapod soon realized the high cost and inefficiency of picking orders in this fashion, and decided to move towards a centralized distribution model in every market in which it offers online shopping and delivery services (Peroni, 2001).

This centralized model employs formats for both large and smaller markets: freestanding warehouse facilities and smaller fast-pick fulfillment centers, depending on the size of its customer base. Through improved supply chain processes, Ahold and Peapod are converting former mezzanine storage areas at Ahold's US stores into efficient "fast pick" fulfillment centers for Peapod.

Peapod works closely with Ahold's operating companies, Stop & Shop in the Boston metropolitan area and Edwards in the New York area. In June 2000, Peapod and Stop & Shop started offering online shopping and delivery service in southern Connecticut. Orders are filled from a dedicated, fast-pick fulfillment center in Norwalk, Connecticut, adjoined to a Super Stop & Shop. The fulfillment center offers a wide variety of Stop & Shop and national brands (Peapod, 2000b).

In Seattle, Albertsons.com uses a retrofitted Albertsons unit and combines a brick-and-mortar operation with a fulfillment center. This strategy employs an existing 31,000-square-foot unit remodeled to present a 14,000-square-foot conventional supermarket with 17,000 square feet reserved as a fulfillment center for picking and packing of orders. Albertsons has continued the strategy of slow expansion using existing stores for order fulfillment as it moved into the Southern California and San Diego markets. As commented on by Matt Muta, Vice President of Technology at Albertsons, "The advantage that we have vs. a centralized fulfillment model (favored by Webvan and other online-only grocers) is that we're not building the multimillion-dollar structures. We are making use of existing structures, existing resources and technologies, and adding the Web front end to it" (Sandoval, 2002b).

In a June 2001 press release, Albertsons Chairman and CEO Larry Johnston discussed his strategic imperatives to be: "... aggressive cost and process control, maximization of return on invested capital, customer-focused approach to growth, company-wide focus on technology, energized associates" (Albertsons, Inc., 2001). These imperatives do not specifically mention the online operations of Albertsons, but the focus on technology, aggressive cost and process control, and maximum return on invested capital facets address Albertsons' approach to its online segment.

With the entrance of large brick-and-mortar grocers into the e-grocery segment, the model of large, automated warehouses has migrated to a larger number of smaller fulfillment centers in conjunction with the existing stores. There are several advantages to the smaller, more dispersed centers. It is often less costly to modify existing stores or structures than to build gargantuan warehouses. Therefore, existing businesses can experiment with the e-grocery segment without a massive outlay of resources. Compared to store-based fulfillment, order picking can be done more efficiently, and it is easier to scale for larger volumes. Having more centers reduces transportation costs due to shorter distances to consumers' homes, and also increases delivery time accuracy and hence customer satisfaction. Additionally, brick-and-mortar grocers that utilize warehouses for online sales as well as traditional distribution to their stores have the potential to increase their efficiency by increasing opportunities for risk pooling, leading to reduced inventory, stock-outs, and lead-times; shared resources, leading to reduced overheard costs; and reduced inbound transportation costs (Beamon, 2001). A problem with having a warehouse handle both online and regular orders to stores is the difficulty of integrating the online and traditional handling, inventory and storage systems. Also, transportation costs to the centers from the suppliers increase because of the additional mileage required. For a detailed discussion on issues and challenges of designing and operating bi-functional distribution centers serving both traditional retail outlets and online orders, see Beamon (2001).

4. Order Delivery

Prior to the advent of supermarkets and chain grocery stores, home delivery of groceries was a common occurrence. Milk was delivered fresh each morning on the doorstep and other groceries were delivered from the corner grocery. Home delivery of groceries has mostly disappeared along with the corner grocery store as huge supermarkets drove small "mom-and-pop" stores out of business. E-grocers are now trying to bring back the home delivery service, together with the customized and personalized shopping experience offered by traditional grocery stores. There are currently four ways by which e-grocers deliver goods to customers: attended delivery, unattended delivery, in-store pickup, and third party pickup locations (see Table 8.2). In this section, we give an overview of each method and discuss their relative advantages and disadvantages. For an in-depth study of attended vs. unattended delivery models in the e-grocery industry, see Punakivi (2003).

4.1 Attended Delivery

Customers of e-grocers, which offer attended home delivery, can usually choose a time window to receive their delivery. In most cases, customers need to place or finalize changes to their orders at least one day before the scheduled delivery time window. For example, FreshDirect allows customers to schedule next day delivery from as late as midnight for weekday deliveries and 9pm on weekend deliveries. Fairprice (in Singapore) makes deliveries 6 days per week with orders received prior to 8am delivered the same day and all others are delivered the following day.

Most e-grocers view having and operating their own delivery fleet of trucks or vans as a strategic advantage. All but two of the top twelve e-grocers ranked by Gomez Advisors Spring 2000 Survey of E-Grocers use company-owned vans or trucks to deliver products to customers. NetGrocer uses FedEx, and GrocerOnline uses both UPS and FedEx for delivering goods to customers.

Table 8.2. A summary of the general logistics model of each company. The level of automation varies and several companies have tried multiple models, including the concurrent use of more than one model.

	Large Automated DC	Smaller, not as highly automated	In-store Picking	Attended Delivery	Unattended Delivery	Installed Receptacles	In-store Pickup
Peapod		**	* (early)	*	*		
PublixDirect	*			*			
NetGrocer	*				*		
Safeway/GroceryWorks	* (early)	* (early)	* (now)	*			
Albertsons		*	*	*			*
Tesco			*	*			
Sainsbury's	* (now)		* (now)	*			*
Peachtree Network			*	*			
Grocery Gateway		*		*			
Fairprice		*		*			
Cold Storage		*		*			
ParkNShop	*			*			
Webvan	*			*	*		
Homegrocer		*		*			
Shoplink		*			*	*	
Streamline		*		*	*		
Kozmo			*	*			

For e-grocers, which own and operate their own delivery network, an important issue is to assign and meet delivery time windows. This requires dynamically assigning delivery time windows to customers as new orders arrive, and dynamically creating and adjusting delivery routes for trucks. High demand for certain time slots, travel time uncertainties due to traffic and other factors, and short time windows further complicate this task. The objectives include maximizing vehicle utilization and minimizing costs, while maintaining acceptable customer service and satisfaction rates. While some e-grocers use commercial routing software, others prefer to use home grown systems or customized versions of commercial software that better meet their needs. Mike Smith, former director of distribution at Homegrocer, noted the challenges in getting routing companies to move at 'Internet speed': "The routing companies are used to working with traditional companies that are not tremendously technology oriented like the dot.coms. If HomeGrocer wants an important change in the software, we often do it ourselves because it is quicker, and we have the technology know-how" (Partyka and Hall, 2000).

One area that differentiates the routing problem faced by e-grocers compared to traditional companies is the need for creating "dynamic" routes which meet short time windows, i.e., creating and adjusting the routes dynamically as orders are placed or changed online. For example, Webvan's windows were 30 minutes long, and Kozmo attempted to form routes within 5 minutes, while taking into account the size and weight of orders, mode of delivery, product type, etc. "Routing companies weren't ready to work on that kind of solution." said Chris Kantarjiev, formerly of the Webvan technology staff (Partyka and Hall, 2000). "Only two vendors showed interest, and we had to customize a commercial package to meet our needs. To make our windows, we take a hit on efficiency and occasionally spend more time driving than delivering." According to Scott Evans, former vice president of Logistics at Kozmo, "We went shopping for routing software last August and decided the offerings were unworkable. They are built for a different type of operation and the algorithms take too long to optimize" (Partyka and Hall, 2000). Kozmo decided to create its own in-house routing system.

To meet the high expectations of on-time deliveries while keeping the delivery costs low, e-grocers need to use fairly advanced optimization techniques and information technology systems. "We have great systems," says Brownell, formerly an industrial engineer at the Webvan distribution center in Suwanee, Georgia (O'Briant, 2000). "When the customer logs onto the Web site and goes to schedule a delivery, we've already done some pre-processing that allows us to know where that customer is. When it shows them the available 30-minute windows that they can schedule, it only shows them the ones that we are able to service them within. So, the system has already done some logic to understand where the customer is, whether there are other orders in the neighborhood, and whether we have capacity available at that given time for that customer to place the order." The routes are built as customers place their orders, and their effectiveness and accuracy are double-checked daily. Brownell adds, "After they place their order, we've held a spot for them in a route. Then at the end of the day, after cutoff, we break all those routes apart; we re-optimize just to make sure that what we've got throughout the daily process is the best match for our utilization of our equipment and for customer service" (O'Briant, 2000).

The concept of integrated order promise and delivery decisions is an interesting one, and to our knowledge, not much studied in the literature (Campbell and Savelsbergh, 2003). While deciding which delivery time windows to offer to a customer, e-grocers need to consider whether (and how) the probability that a customer places an order depends on the available time windows. For example, if the only delivery times offered to a particular customer are between 2pm and 6pm, would the customer still place an order? E-grocers should also determine how this customer's delivery request fits into the current delivery schedule, given current and potential future orders. Of particular concern are the potential arrivals of more profitable orders in the future, which could utilize the time slots offered to the current customer. With the goal of creating balanced delivery schedules that lead to an efficient and effective utilization of the existing fulfillment/delivery capacity, we envision that revenue management techniques could be applied to the delivery fees. For example, e-grocers could set different delivery fees depending on the delivery time and day, similar to package delivery services; e.g., higher fees for popular time windows or for morning or evening rush-hour times. Delivery fees could also be dynamically adjusted depending on how a new order fits into the current delivery schedule. For example, the delivery fee can be waived (or lowered) if a customer accepts her delivery in the same window as another customer (who already placed an order) from the same neighborhood. Webvan had taken a first step in this direction by highlighting the time windows at which deliveries were already scheduled to a customer's neighborhood. However, since Webvan did not offer any discounts or other incentives for choosing these particular windows, it is not clear how much that information influenced the customers' decisions.

Creating dynamic routes given tight delivery windows and uncertainties in demand and travel times is an extremely difficult task. To increase delivery efficiency and timeliness, some companies put restrictions on delivery times. For example, FreshDirect, which operates in New York City, limited delivery slots to evenings (4-10pm) to minimize delays due to traffic. Concentrating delivery times to this shorter time frame has a two-fold benefit for the company. First, the trucks avoid morning rush hour, minimizing the time they are immobilized in traffic. Second, the shorter delivery time frame concentrates the demand per window, leading to fewer trips, better truck (and driver) utilization, less gas usage, and lower wear and tear on the vehicles.

Besides the debates on the profitability of the home delivery model, the views are also divided about the impact of this model on the environment. Some opponents of the e-grocery model cite state that e-grocers do more damage to the environment than traditional stores. Their studies show that fuel emissions from operating home delivery services are higher than the combined trips of individuals shopping for groceries (Galea and Walton, 1997). Although environmental issues are a concern, we do not discuss them further in this chapter.

4.2 Unattended Delivery

Currently, only two e-grocers, Netgrocer and Peapod, are offering unattended delivery. Netgrocer delivers only non-perishables via FedEx. Peapod drivers leave insulated coolers packed with dry ice in a customer-designated secure location, and pick the coolers up on their next delivery. Peapod's method of unattended delivery is fairly low cost (coolers and dry ice) and the benefits gained by customer goodwill due to the added convenience may outweigh the cost. Using a national carrier for delivery, as in the case of Netgrocer, takes off this burden from the shoulders of the e-grocer, but increases the possibility of spoilage and theft.

Previously, e-grocers used other methods for unattended delivery. For example, Shoplink.com and Streamline.com installed refrigerated boxes or storage shelves at customers' garages to hold the groceries. Streamline.com provided two delivery options: orders could be delivered in specially designed temperature-controlled bins, or placed in a full-sized Streamline.com refrigerator. The refrigeration unit, which Streamline.com designed with Sears, sat primarily in a garage or basement area for each weekly delivery. Streamline installed a keypad entry system or lockbox to the customer's garage or basement to be able to make the delivery while the customer was not at home.

According to a recent study, unattended delivery is found to be the most cost-efficient e-grocery home delivery model, since it enables the optimal routing and scheduling of delivery vehicles (Punakivi, 2003). Unattended delivery also appeals to the lifestyles of the majority of e-grocer customers, who are working individuals not likely to be at home during the day to receive deliveries (Blackwell, 2001). Despite its cost advantages and convenience for the customers, this method has serious drawbacks. First, apartment dwellers cannot partake of this service unless packages can be left at a management office or with a doorman. Second, even if the customers do not pay for the storage boxes, they still have to give up part of their precious garage or storage space. Third, the cost of buying and installing storage boxes and shelves is relatively high and has to be amortized across the expected duration of the boxes' use. Punakivi (2003) finds that while using customer-specific reception boxes in home delivery operations leads to a cost reduction of 44-53 percent (compared to attended delivery with a two-hour time window), due to the high investments involved in customer-specific storage boxes, the payback period is 6-13 years.

Although unattended delivery may be an attractive service for some customers, the reality of its implementation is prohibitive or undesirable for most e-grocers. Finding a method of keeping perishable grocers cool and safe until customers come home and store them away, can be difficult, or costly, or just plain time-consuming. With the other available methods of getting the goods to the customer on the go, such as in-store pick-up and third-party pickup locations, unattended delivery is a serviceable, but not necessary option.

4.3 In-Store Pickup

With so many traditional supermarket chains joining the ranks of e-grocers, they leverage their assets for yet another feature they possess not available to pure-plays - in-store pickup. Albertsons.com is one of the online grocers currently offering this service. Especially for those customers constantly on the go, or unsure about scheduling, this may be a desirable option. The downside to in-store pickup is that customers still have to make a trip to the grocery store. However, with lower delivery costs for the e-grocer and the customer, this method certainly has a potential to be around for much longer.

4.4 Third Party Pickup Locations

Customer pickup in third party locations has been long discussed in the e-grocer industry, but none of the e-grocers in the U.S. provided this option until recently. In Japan, convenience stores have been used for some time for picking up orders placed online (Strom, 2000). Similarly, in Taiwan there have been partnerships between online stores and convenience stores for payment and delivery. For example, the partnership that took place between Music.com.tw and FamilyMart in 1998 allowed consumers order music CDs online, and then pay and pick up the CD at any of FamilyMart's outlets. More than 6,000 CDs were sold in just a few weeks (Ling, 2000).

FreshDirect is now using office parks and train stations as remote pickup locations for groceries in suburban New York. According to FreshDirect CEO, Joseph Fedele, refrigerated trucks carrying up to 500 customers' groceries wait at designated locations from 2-8pm for customer pickup. This adds to the convenience factor for on the go customers by having someone else waiting for deliveries, says Fedele, "You don't wait for us to come. Our trucks wait for you" (O'Connell, 2002). An interesting question is how to choose such pickup locations. In these decisions, e-grocers need to tradeoff convenience to the customers against the cost. For example, convenience to the customers could require a higher number of pickup locations possibly in areas where land is expensive and congestion is high, which would increase the costs. The rich literature on facility location problems (Drezner, 1995) could provide useful insights to e-grocers in evaluating alternative pickup locations and choosing the best ones.

If the customer pickup method catches on, we may see in the future an expansion of this service, perhaps placing refrigerated containers with personalized locks in mall parking lots, gas stations or convenience stores. E-grocers as well as other retailers could form alliances to lower costs by delivering through third-party or shared facilities. An alliance could consolidate customers' orders from multiple e-tailers in "collection centers" built in convenient locations, which could be a new facility or perhaps a gas station, convenience store, or a local Wal-Mart store, and then either deliver from there, or let the customers pick up all the goods they ordered online (see Figure 8.4). Consolidating online orders from multiple retailers could result in reduction of costs of delivery as expenses are spread between several companies.

5. Conclusions

Just a few short years ago money seemed to be plentiful. Almost anyone with a good idea could obtain venture capital, start a business, and if it failed it was considered a "learning experience." Those days are gone. The rules of old business now apply to 21st century businesses, especially e-businesses. Companies must not only have a good idea, but they must have competent management, solid logistics, and must be profitable to stay relevant and stay alive. This rule applies to all companies, regardless of sector, but the logistics aspect is particularly crucial

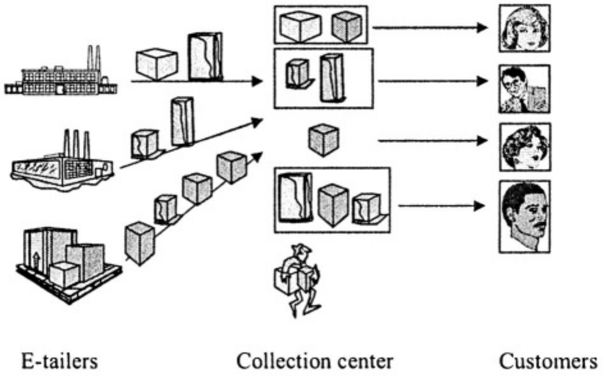


Figure 8.4. Collection centers consolidate the orders a customer placed from multiple retailers, for one-stop pickup or aggregate home delivery.

for e-grocers. Grocery business operates on thin margins, and hence it is imperative that e-grocers identify the most efficient and effective ways for order fulfillment and delivery of the bulky and/or perishable items they sell.

Today, the surviving and thriving online grocers are not pure-plays but companies such as Safeway, Albertsons, and Publix. The existing giants' most predominant advantages over pure-plays are massive resources, logistics infrastructures, and existing customer base, which lower the barrier to entry of traditional grocery stores while maintaining the high barrier to entry for pure-play e-grocers. Current trends indicate that survival is quite difficult, if not impossible, for e-grocers with no "brick base," with the notable exception of niche players. Albertsons, Safeway, Publix, and other established grocers with online stores will grow stronger, especially in those cities with dense populations where driving to a grocery store is inconvenient and the cost of groceries is high (New York, Chicago, San Francisco, etc.). Once consumers have more than one online grocery service in each market to choose from, customer loyalty will be a key factor in an e-grocer's success. One study shows that "product availability, timeliness of delivery, and ease of return have statistically significant association with customer loyalty" (Heim and Sinha, 2001). Customer-acquisition costs for a startup online grocer range between $200 and $700 per customer (Kumar, 2001). This means that attracting consumers alone will not determine success, but retaining existing customers is crucial.

One mistake most e-grocers made in facing the challenges of this business was rapid growth, both in terms of the markets served and the variety of product/service offerings. Back in the late 1990s, most e-grocers, as well as other e-tailers, went by the philosophy "If you build it, they will come." They spent millions of dollars on state-of-the-art warehouse facilities, fleets of trucks, and advertising, but could not build the customer base needed for break-even, let alone profitability. Most customers were not ready to change their shopping habits, and even if they were, they preferred to buy from existing stores selling online, which they found familiar and more reliable.

Similar to rapid expansion into too many markets, rapid expansion into too many products and/or services also proved to be costly for most e-grocers, especially pure-plays. For example, now defunct Streamline.com's services included fresh flower delivery, shoe repair, dry cleaning, UPS package pickup and mailing, film development, and video rental. Further along those lines, some e-grocers, such as Webvan, wanted to position themselves less as a grocery player and more as a fulfillment player. Webvan repeatedly stressed that the notoriously low-profit-margin grocery business (the big chains earn about a 2 percent margin) is not its main business. Once grocery markets were established and a strong brand image was built, Webvan's plan was to move into higher margin products. That is one reason why the company originally choose the name Webvan, rather than Webgrocer or some other name that directly associates it with the grocery industry. Instead, Webvan wanted to be a "last mile" Internet retailer, the retailing equivalent of telecom's "last mile" companies - the firms that want to own that critical last-mile connection to customers' homes and pockets. Webvan would then compete as much with UPS for rapid package delivery as it did with Safeway for food sales. The attraction of such home delivery fulfillment services is that they have an opportunity to prompt a lot more impulse purchases than the Internet normally does.

In many ways, such a move set e-grocers like Webvan and Streamline.com on a collision course with other retailers, such as Amazon.com, who were moving from being a specialist towards becoming a generalist. This attempted product/service diversification strategy of e-grocers was partly due to their need for finding ways to generate new revenue sources, increase margins, and better utilize existing resources, in particular, the capacity of their highly automated warehouses and delivery networks. That is why Amazon took stakes in HomeGrocer and Kozmo.com, companies that bypassed postal services like Federal Express and UPS to deliver products right to consumers' homes.

On the other side of the spectrum, there are those companies that take pride in catering to a smaller, select group of customers. Already we see New York's FreshDirect focusing on fresh, perishable foods. CEO

Fedele seems to distance his company from other e-grocers and online retailers, stating "This isn't about the Internet. It's about offering a better, fresher product at prices 10 percent to 35 percent cheaper ... We understand the money is made in expert manufacturing, not in distribution" (Dillon, 2002). With the setup at the FreshDirect processing center, one might opine that they could broaden their scope to delivery of fresh products to restaurants, hotels and specialty stores, themselves becoming middlemen, should the online delivery business not progress as anticipated; however, nothing along those lines has been stated publicly by the company. To avoid the Webvan mistake of overreaching, FreshDirect delivers in only five zip codes, mostly in Manhattan, and is adding new ones slowly as it fine-tunes its systems. It has no immediate plans to expand beyond New York; the longer-term goal is to open one or more additional plants in the region before expanding to four or five other East Coast cities (Kirkpatrick, 2002).

Currently, the existing e-grocers in the U.S. and Asia are not extending their product lines to include higher margin items, such as DVDs or dry cleaning services. They are focusing on their core business of groceries/consumer items and using home delivery as an alternate channel, rather than expanding into items outside their current product mix. In Europe, however, the trend is mixed, with some retailers carrying high margin, non-grocery items, and others are not. For example, U.K.'s Tesco expanded its product selection beyond groceries to include higher margin products that online customers seem more willing to mix with food than do customers in the supermarket. We have yet to see the full of effect of Tesco's strategy but it may well turn out that the answer may be a careful selection of additional high-margin products and, perhaps most importantly, a familiar store brand from which to buy all these goods.

There is room for many niche suppliers in the online world. We see opportunity for growth in the niche areas of fresh and prepared foods (e.g., FreshDirect), non-perishable pantry items (e.g., NetGrocer), convenience store items, and alcoholic beverages. There is also potential in the ethnic markets. Certain companies such as EthnicGrocer.com and LatinGrocer.com are already taking advantage of this opportunity by catering to the Hispanic market, but there are yet untapped markets to be reached in the future.

A number of consumers, investors, and traditional retailers remain skeptical of the promise of online grocery retailing. At the heart of all the criticism is that the grocery business is mature and has very thin margins, which are not high enough to offset the high costs of order fulfillment and home delivery for online orders. "Running a delivery service

- with trucks and staff to maintain - is pricey." Says Ken Cassar, a senior analyst for Jupiter Research: "It is a very expensive business to bid out, and consumer habits die hard. Consumers have been accustomed to selecting their own tomatoes, and they don't trust anyone else to pick them out for them" (Anonymous, 2000a; Mott, 2000). Yet we see opportunity for growth and success for a select few online grocers. To summarize, our recipe for success in the online grocer business is as follows: don't grow too soon, too fast; partner with existing brick-and-mortar companies; design logistics infrastructure for efficiency and scalability; attract and retain customers; communicate the value clearly; and do not try to compete on price, but on overall "value." If online grocers combine these ingredients we expect the results will be highly palatable.

Acknowledgements

Pınar Keskinocak is supported by NSF Career Award DMI-0093844. This research is also supported in part by a grant from the Logistics Institute Asia-Pacific (TLI-AP).

References

Ahold (2002a). Ahold third quarter 2002 media information. www.ahold.com/mediainformation/news/article.asp?news_id=458.

Ahold (2002b). Royal Ahold 2002 annual report. www.ahold.com.

Albertsons, Inc. (2001). Albertsons, inc., exceeds first quarter 2001 earnings expectations - defines focused strategic imperatives to drive sales and earnings growth. Albertsons Press Release.

Albertsons, Inc. (2002). Albertsons.com provides new grocery shopping alternatives to Bay Area residents. Albertsons Press Release, www1.albertsons.com/corporate/default_news.asp?Action=Continue&ContentId=1301.

Anonymous (2000a). Part one: We will pay more attention to customers. *Business 2.0,* 5:83–84.

Anonymous (2000b). Shopping around the web: A survey of e-commerce. *The Economist,* 354:5–54.

Anonymous (2001). British grocer get 35 percent stake in GroceryWorks.com. *Dallas Business Journal,* http://dallas.bizjournals.com/dallas/stories/2001/06/25/daily10.html?t=printable.

Bartholdi, III, J.J. and Hackman, S.T. (2003). Warehouse and distribution science, www.tli.gatech.edu/whscience/book/wh-sci.pdf.

Beamon, B.M. (2001). Multiple order-entry points in e-business: Issues and challenges. In *Proceedings of the 2001 IEEE International Engineering Management Conference.*

Beck, E. (2000). British grocer Tesco thrives filling web orders from its stores' aisles, October 16. *The Wall Street Journal – Eastern Edition,* 236(74):B1–B13.

Bellantoni, C. (2000). Webvan tops fourth quarter IPOs in valley. *Silicon Valley/San Jose Business Journal,* http://sanjose.bizjournals.com/sanjose/stories/2000/01/17/story3.html.

Blackwell, R. (2001). Why Webvan went bust. *The Wall Street Journal – Eastern Edition,* July 16, 238(10):A22.

Callahan, D. (2000). Editor's view: Streamlining the e-tailers. *Grocery Central,* www.grocerycentral.com/, September 18.

Campbell, A.M. and Savelsbergh, M. (2003). Decision support for consumer-direct grocery initiatives. Technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology.

Davis,P. (2001). Sainsbury 2001 annual report. www.j-sainsbury.co.uk/anrpt01/howwedid/grpchief.htm.

Desrosiers, J., Dumas, Y., Solomon, M., and Soumis, F. (1995). Time constrained routing and scheduling. In Ball, M., Magnanti, T.L., Monma, C.L., and Nemhauser, G.L., editors, *Handbooks in Operations Research and Management Science - Network Routing,* pages 35–139. North-Holland, Amsterdam.

Dillon, N. (2002). Fairway grocer bucking odds with web service. *New York Daily News,* www.freshdirect.com/about/press/article.jsp?catId=about_press_recent&articleIndex=2&trk=snav.

Drezner, Z., editor (1995). *Facility Location: A Survey of Applications and Methods.* Springer-Verlag.

Fabricant, F. (2002). Fresh groceries right off the assembly line. *The New York Times,* www.freshdirect.com/about/press/article.jsp?catId=about_press_recent&articleIndex=3&trk=snav.

Galea, C. and Walton, S. (1997). *Is E-commerce Sustainable? An Analysis of Webvan, Sustainability in the Digital Economy.* Greenleaf Publishing.

GroceryWorks (2003). www.groceryworks.com.

Heim, G.R. and Sinha, K.K. (2001). Operational drivers of customer loyalty in electronic retailing: An empirical analysis of electronic food retailers. *Manufacturing and Service Operations Management,* 3(3):264–271.

Hodges, J. (2000). Jumping on the brandwagon: Online upstarts know how to size up nascent markets. *Business 2.0,* January 1.

Kahn, B. and McAlister, L. (1997). *The Grocery Revolution: The New Focus on the Consumer.* Addison-Wesley, Reading, Massachusetts.

Kamarainen, V. (2003). *The impact of investments on e-grocery logistics operations.* PhD thesis, Helsinki University of Technology, Industrial Engineering and Management, Helsinki, Finland.

Kane, M. (2002). Tesco aims where Webvan flamed. *CNet News,* `http://news.com.com/2100-1017_3-813011.html`, January 14.

King, C. (2001). Webvan reroutes business model. `www.internetnews.com/ec-news/article.php/570301`, January 25.

Kirkpatrick, D. (2002). The online grocer version 2.0. *Fortune,* November 25, 146:217–221.

Knowledge at Wharton (2001). What makes a winning net grocer. *CNet News,* `http://news.com.com/2009-1017-274385.html`, October 20.

Kumar, N. (2001). Online grocers take stock. *Business 2.0 (U.K. Edition),* March 29, pages 46–52.

Laporte, G. (1992). The vehicle routing problem: An overview of exact and approximate algorithm. *European Journal of Operational Research,* 59:345–358.

Lee, L. (2003). Online grocers: Finally delivering the lettuce, brick-and-mortar chains are finding profits in cyberspace. *Business Week,* April 28, 3830:67.

Ling, C. (2000). Convenience stores in Taiwan become ecommerce destinations. *The Wall Street Journal Interactive,* April 19.

Maddali, P. (2003). Tesco in 2003 case study. ICFAI Knowledge Center, Reference No: 15-03-06-08, `www.icfai.org/ikc/main/Articles/padmini/tesco.PDF`.

Millenium (2003). Tesco.com: A Millenium Group case study. `www.millenn.co.uk/`.

Moore, J.F. (2001). Why Peapod is thriving: First-failure advantage. *Business 2.0,* August 14.

Mott, S. (2000). Resolutions for 2001: What you can learn from a year of hell. *Business 2.0,* pages 78–97.

Muehlbauer, J. (2001). Webvan gets totaled. *The Industry Standard,* July 10.

O'Briant, E. (2000). Webvan revs up. *IIE Solutions,* pages 26–33.

O'Connell, P. (2002). Can FreshDirect bring home the bacon. *Businessweek Online,* Small Business Entrepreneur Q&A, `www.freshdirect.com/about/press/article.jsp?catId=about_press_recent&articleIndex=4&trk=snav`, September 24.

Partyka, J.G. and Hall, R.W. (2000). On the road to service. *OR/MS Today,* August, pages 26–30.

Peapod (2000a). Ahold forms partnership with us internet grocer peapod. `www.peapod.com`, April 14.

Peapod (2000b). Peapod by Stop & Shop offers online grocery shopping and delivery service in Connecticut-model illustrates benefits achieved with Peapod/Royal Ahold partnership. www.peapod.com.

Peroni, P. (2001). Two grocery e-tailers to exit markets. www.pfma.org/media/advisor/MAR01/columns/CentralIssues.html, March.

Punakivi, M. (2003). *Comparing alternative home delivery models for e-grocery business.* PhD thesis, Helsinki University of Technology, Industrial Engineering and Management, Helsinki, Finland.

Reinhard, A. (2001). Tesco bets small - and wins big. *Businessweek Online,* www.businessweek.com/magazine/content/01_40/b3751622.htm, October 1.

Richtel, M. (1999). For new online grocer, high-tech warehouse is key. *The New York Times,* June 3.

Rubin, R. (2001). Online grocerys second wind. www.forrester.com/ER/Research/Report/Summary/0,1338,11542,00.html.

Sandoval, G. (2002a). Grocers make another go at home delivery. *CNet News,* http://news.com.com/2100-1017-827899.html, February 1.

Sandoval, G. (2002b). Net supermarkets next wave. *CNet News,* http://news.com.com/2008-1082-852107.html, March 5.

Smaros, J., Holmström, J., and Kämäräinen, V. (2000). New service opportunities in the e-grocery business. *International Journal of Logistics Management,* 11(1):61–74.

Sneader, K., Crawford, B., and Ebrahim, S. (2000). Operation food: Fighting to win in the new economy. www.fmi.org/e_business/presentations/OperationFood/sld001.htm, June 26.

Strom, S. (2000). E-commerce the Japanese way. *The New York Times,* March 18, pages C1–C4.

Tapscott, D. and Ticoll, D. (2000). Retail evolution. *The Industry Standard,* July 17.

Wallace, P. (2000). Wicked Amazon wannabe? European grocer hopes to make it big on the web. *The Industry Standard,* June 8.

Yrjölä, H. (2003). *Supply Chain Considerations for Electronic Grocery Shopping.* PhD thesis, Helsinki University of Technology, Industrial Management and Work and Organisational Psychology, Dissertation Series No.3, HUT, Helsinki, Finland.

Chapter 9

ENABLING SUPPLY-CHAIN COORDINATION: LEVERAGING LEGACY SOURCES FOR RICH DECISION SUPPORT

Joachim Hammer
Department of Computer & Information Science & Engineering
University of Florida
jhammer@cise.ufl.edu

William O'Brien
M.E. Rinker, Sr. School of Building Construction
University of Florida
wjob@ufl.edu

Abstract As supply chains come under escalating pressure for responsiveness to customer demands as well as cost pressures, there is increased scope for decision support models to enhance coordination of supply chain activities. However, enhanced coordination implicitly requires sharing information about costs, production capacity, materials availability, delivery schedules, etc. This information is currently stored in a panoply of legacy information systems distributed across the many firms that comprise a given supply chain. As such, accessing useful information presents considerable technical challenges, effectively limiting the deployment of decision support tools for supply chain operations. The purpose of this chapter is to present the research issues and approaches to accessing or integrating data in legacy information systems. Current industry approaches to information integration, including information hubs and Web Services technologies, are briefly reviewed and critiqued as useful but inadequate to the task of integrating data stored in legacy systems. Recommended requirements for new methods include rapid deployment, ability to connect to heterogeneous legacy systems, composition of knowledge for decision support, and provision for secure data access. These requirements motivate a review of the research literature on knowledge extraction and composition. As an example of new methods built from current research, an integrated toolkit known as SEEK: Scalable Extraction of Enterprise Knowledge is presented. Capabilities and limitations of the SEEK toolkit are used to suggest novel areas of research in visualization and representation of data for human refinement of automatic integration results, as well as further development of evolutionary algorithms to enhance the scope of automatic knowledge extraction. Throughout the chapter, an example of a construction industry supply chain is used to motivate discussion.

1. Introduction

Firms in a supply chain must share at least a base level of information to coordinate their actions. Much current thinking in both industry and academia suggests that improved sharing of information can lead to improved coordination for faster response, increased flexibility, and lowered inventory, transport, and production costs (Lee and Whang 2002; Tan et al. 2000). The question is how to achieve these benefits? Organizational barriers to sharing information can be enormous, and there is a growing literature in the design of contracts to promote sharing of information and incentives for system response (Tsay et al. 1999). But even given the organizational will to collaborate, firms in a supply chain face enormous technical hurdles in integrating their information systems. Despite the promise of the Internet, XML based standards for data exchange, and Web services architectures such as Microsoft's .NET™, it remains a time consuming and expensive task for firms to share information or access sophisticated supply chain decision support tools. In short, the heterogeneity present in the types and sophistication of firms that compose a supply chain, and the associated heterogeneity in those firms' information systems, presents a structural challenge that cannot be overcome with a few data standards. There is a need to generate new methods that provide semi-automatic approaches to leverage firms' knowledge for rich decision support. These methods will provide the basis to enable e-commerce and supply chain management paradigms based on rich and rapid sharing of firms' information.

This chapter serves as both a primer and research discussion about approaches to leverage the knowledge resident in supply chain legacy sources. By legacy sources we mean computer applications, databases, and associated hardware resident in a firm. As such, our use of the term legacy source includes current applications based on modern specifications as well as proprietary systems. Section 2 relates supply chain coordination approaches to the data and knowledge resident in legacy sources, concluding with a list of requirements for new systems and methods. Section 3 discusses the main approaches to extracting legacy knowledge and composing that knowledge for subsequent use by supply chain decision support tools and decision makers. Section 4 reviews an integrated toolkit - SEEK: Scalable Extraction of Enterprise Knowledge - for supporting rapid integration of legacy sources with decision support tools (O'Brien et al 2002). Section 4 provides an architectural discussion of the SEEK approach as a structured, secure approach to knowledge extraction as well as a detailed example. Building from the authors' experience with SEEK, Section 5 suggests research needs to radically enhance the capabilities of data extraction and composition tools. Some concluding remarks are made in Section 6.

2. Supply Chain Coordination: Relationship to Legacy Sources

Past research in the operations management literature widely recognizes that information sharing can potentially lead to improvements in coordination among firms in a supply chain. Several models and decision support tools have been developed that can help firms better coordinate their operations with shared knowledge of demand, inventory levels, production capacity, etc. (e.g., Beyer and Ward 2002; Gallego and Özer 2002). Concomitant with the advent of such tools have been calls for information technology (IT) applications to support supply chain coordination. The advent of the Web has in particular been used as a rallying point for calls for cheaper, faster, and more capable IT solutions. However, a fundamental barrier to effective implementation has been the ability to utilize firms' legacy information for supply chain decision support applications.

As a motivating example, consider a construction general contractor that requires frequent updating of resource availability from its subcontractors. The process of updating the master schedule is currently performed manually, with associated information collection difficulties and limited ability to perform timely resource constrained optimization (O'Brien et al. 1995). Sharing electronic information across the construction supply chain is difficult as subcontractors have diverse legacy information systems with different levels of detail (Castro-Raventós 2002). For example, a contractor may maintain an aggregated master schedule whereas a subcontractor will break down a specific activity such as "erect steel area A" into a detailed erection sequence. Hence, collection of the information necessary to update or compute an optimal master schedule (an information hub type application as described in Section 2.1) requires considerable effort to link to data sources and transform detailed information into the aggregated form appropriate for the master schedule. Existing manually instantiated computing solutions do not scale to the needs of the construction supply chain. There are too many subcontractors (with an associated panoply of information systems) to manually build links between each subcontractor and the general contractor's master schedule. There are also significant issues of information privacy and security with manual creation of links (an issue addressed in Section 4.1). The inability to access firms' legacy information retards effective deployment of supply chain coordination tools, with associated problems in practice (e.g., poorly coordinated schedules are largely responsible for the low workflow reliability of 30-60% observed in construction (Ballard 1999)).

2.1 Hubs and related paradigms for supply chain coordination

Supply chain coordination on construction projects is a particularly pernicious example of the difficulties of leveraging legacy information to support supply chain coordination. However, the problems faced by construction companies are by no means unique. Several authors have proposed e-commerce solutions to support supply chain coordination issues. Tan et al. (2000) catalog supply chain processes supported by Web applications and anticipated benefits. Tan et al. (2000) also recommend that Web based technologies to support supply chains be built from *component-based* technologies to support modularity and flexible modeling of supply chain processes. A somewhat different perspective is given by Lee and Whang (2001; 2002), who advocate an *information-hub* based architecture (Figure 9.1) to support supply chain coordination. The hub acts as an information broker and value added tool to support supply chain operations. Perhaps operated by a third party serving as application service provider (ASP), the hub also provides security by shielding a firm's data from other firms in the supply chain. At the same time, persistent connection to a hub also provides for continuous monitoring of supply chain operations and hence acts as an early warning system for potential problems. Thus Lee and Whang (2001; 2002) note that a certain level of trust among firms is needed for hubs to operate effectively. An alternate approach to a centralized, hub-based architecture is implementation of an agent-based system for decentralized coordination (Verdicchio and Colombetti 2002). Under this approach, each firm in a supply chain employs one or more agents to gather data or to make/report commitments with other firms.

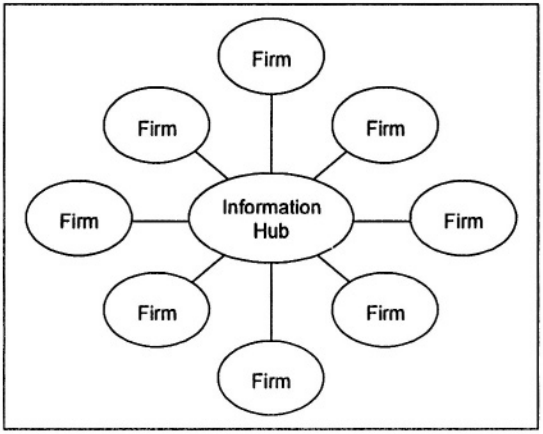


Figure 9.1. Hub-based architecture for supply chain coordination.

The basic limitation of the component, hub, and agent-based approaches are that they put the burden of integration on the firm. It is very easy to draw a line showing a connection on paper. However, implementation of either the hub or the agent architectures requires a queryable interface to the firm's legacy sources. This is a non-trivial task. Shapiro (1999; 2001) identifies ten types of decision support applications (including MRP and APS systems) that individual firms may implement for production control and supply chain modeling. In addition, data relevant to supply chain decision support may be stored in product catalogs, price lists, accounting systems, etc. Queries from a hub such as "are resources available for product x in the period between May and June?" may require data from the MRP database, the product catalog, and a database of labor resources. Building a tool that sits across these databases and answers queries from agents or hubs is an expensive proposition even for the largest firms.

A component-based approach faces similar challenges of integration. A suite of components may integrate well with each other but still must be integrated with the numerous internal systems of a firm that are not part of the suite. And should those components contain data that must interact with components or hubs or agents built from differing standards, we return to the same problem of multiple systems that must be queried. Hence implementation of decision support tools that use data from multiple firms presents a fundamental difficulty of systems integration, as shown in Figure 9.2. At the moment, the burden for integration is placed on the firm, making each implementation unique.

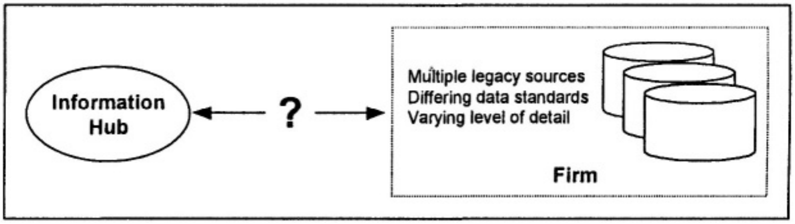


Figure 9.2. Integrating legacy sources with external systems: how to bridge the gap?

Supply chain decision support solutions are thus difficult to scale to large applications across multiple firms.

2.2 Are data standards a panacea?

At this point the reader may ask, but what about XML based data standards? Are they not designed to overcome these issues? The authors' experience to date indicates that data standards will help, but are by no means a panacea. With a large supply chain, it is unlikely that all the firms involved will subscribe to a common data standard. For example, a large manufacturer of high-value, small-volume products in south Florida maintains a database of 3,000 suppliers, 1,000 of which are active at any one time. Given the range of sizes, sophistication, and disciplines of these suppliers, it is highly unlikely that a single standard will be used. Rather, if firms are to adopt a data standard they are likely to do so for certain disciplines and for certain clusters of firms that find value in repeated interactions. (For discussion of the use and evolution of data standards in the construction industry - known for many small-to-medium size companies that operate in flexible supply chain configurations - see Amor and Faraj 2001, and Zamanian and Pittman 1999.) Thus data standards may help small groups of firms but are unlikely to effectively support the operation of larger supply chains.

Data standards are also limiting as the complexity of information to be shared increases. While in theory data standards should resolve issues with complexity, in practice the more complex the data exchanged, the greater the burden placed on the firm to integrate its existing systems with the proposed standard. Firms may prefer to retain the use of proprietary legacy sources for many if not all operations. Hence, asking those firms to link their existing systems with a data protocol that can be queried by a hub or agent requires significant effort by the firm. Further, even if firms deploy applications built on a common data standard, this does not mean they will uniformly deploy those applications.

Firms have different operating procedures and different levels and types of data collected for internal management. (See Castro-Raventós 2002, for a comparison of different information system implementations among medium size firms.) Thus, even if a common data standard is employed, sharing complex data can require considerable effort to transform from a firm's internal representation to the requirements of the information hub or agent.

There is a need to build translators to help firms bridge the gap depicted in Figure 9.2. To a limited extent, the push to develop Web services such as Microsoft's .NET™ or Apache's AXIS™ provides basic tools to help firms configure links between their data and applications. These tools make it easier for experts; it is likely there will still be extensive manual work required for configuration. Further, .NET™ and similar toolkits are designed for generic business services and it is unclear that they will develop the necessary sophistication to address the complexity of firms' internal data systems for more than basic supply chain integration.

The limitations of generic tools for translation of complex process or supply chain data have been anticipated by standards organizations. Recently, the Process Specification Language (PSL) (`www.mel.nist.gov/psl`) (Schlenoff et al. 2000) and the XML Process Definition Language (XPDL)(`www.wfmc.org/standards/docs.htm`)(Workflow Management Coalition 2002) have been promoted as standards for the exchange of process models and process information. PSL and XPDL are envisaged as neutral formats for the exchange of process information. PSL in particular is envisaged as a kind of process lingua franca that reduces the overall number of translators between heterogeneous tools. However, as noted by Schlenoff et al. (2000), PSL is not a "characterization language" that describes the behaviors and capabilities of a process independent of any specific application. Tools to reason about connections between processes are not native to the PSL or XPDL specifications. Missing from both languages are constructive methods to actively build translators. Hence, implementation of a translator between a given legacy implementation and the specification requires extensive manual effort by experts.

2.3 Requirements for new approaches

Current tools and initiatives, such as XML based data standards, and specifications, such as XPDL, are making it easier for firms to share information and improve coordination across the supply chain. A fundamental barrier to rich data exchange, however, is that the burden of integration remains with each firm. Thus, if fifty firms comprise a supply chain and they wish to use a hub for coordination, there are likely fifty unique integration efforts that must take place. And if some of those firms wish to participate in a different supply chain that uses a different hub or method of coordination, they will likely need to undertake a new integration effort. This is an untenable situation. To balance the needs for data exchange for rich decision support, there need to be semi-automatic methods to construct rapid links between firms and hubs. We suggest that such methods have the following minimum requirements:

- *Rapid deployment:* Supply chains comprise a large number of firms and, with calls for flexibility, are increasingly assembled and disassembled quickly. Systems must be deployed swiftly with limited human interaction.

- *Connect to heterogeneous sources:* The large number of firms in a supply chain suggests that associated legacy sources will not subscribe to uniform data standards, but rather present a high degree of heterogeneity both physically and semantically. A system must accept a wide range of source types.

- *Composition of data:* Concomitant with heterogeneous information representation, there is a need to compose data stored in a firm's legacy sources to support the information needs of supply chain decision support applications.

- *Security:* Firms will generally be unwilling to make their legacy systems open for general examination by other firms. Systems must filter data extracted from the underlying sources.

These functional requirements necessitate use of automatic approaches to the extraction and composition of data, which we next discuss in Section 3. An integrated approach to implementing the extraction and composition methods developed in reference to these functional requirements is then detailed in Section 4.

3. Approaches to Leveraging Legacy Sources

Given a wide range of heterogeneous legacy sources in the supply chain, there are dual challenges in both extracting information and composing that information for subsequent use. And while it is possible to manually construct linkages between a given firm's legacy sources and decision support applications, heterogeneity implies that there is little possibility for reuse of code developed for specific implementations. Hence there is a need for approaches that semi-automatically extract and compose legacy source knowledge for subsequent reuse by supply chain applications. Such semi-automatic approaches will support the coordination problems faced by the construction supply chain as detailed in Section 2. A construction scheduling example is used throughout the chapter; however, the discussion and presentation of techniques is applicable to knowledge extraction and composition in general.

3.1 Approaches to Knowledge Extraction

The state-of-the-art in knowledge extraction includes three major areas of research: data reverse engineering, program analysis, and data mining. This section summarizes the past and current research in these areas. However, a few definitions are in order before beginning our review. In the computer science literature, knowledge discovery and extraction are used interchangeably to refer to the process of mining knowledge from large amounts of mostly unknown data (Han and Kamber 2001). Since it is generally accepted that extraction assumes the discovery of knowledge, we use the term *extraction* exclusively when referring to both the discovery and extraction of knowledge. Further, the term knowledge can mean many things to different people. In the context of this chapter, we use *knowledge* to refer to the domain-specific information and procedures governing the operations of an enterprise. Specifically, we refer to the metadata describing the informational resources of an enterprise such as databases, their semantics, as well as any business rules for the enterprise, which may be encoded in the information itself or in the applications manipulating the information.

3.1.1 Data Reverse Engineering. Software reverse engineering (SRE), a well-known practice in computer science, has as its goal the inference of program structure and functionality from compiled source code. Similarly, *data reverse engineering* (DRE) refers to the inference of structure and meaning (e.g., schema, relations, and semantics) from databases. As SRE is primarily applied to legacy source code, DRE techniques have been applied to legacy source data. For example, Aiken (1996) aptly defines DRE as "the use of structured techniques to reconstitute the data assets of an existing system." The emphasis on structured techniques is key to economically feasible DRE implementations, which attempt to add value to existing data assets by increasing their consistency and ease of use within or across organizations (Davis and Aiken 2000).

Industrial legacy database applications (LDAs) often evolve over several generations of developers, have hundreds of thousands of lines of associated application code, and maintain vast amounts of data. In many cases, the documentation has become obsolete and the original developers have left the project. For our purposes, a major task of DRE is recovery of LDA conceptual structure that is often based on a relational database. Unfortunately, the simplicity of the relational model does not support direct description of the underlying semantics, nor does it support inheritance, aggregation, n-ary relationships, or time dependencies including design modification history. However, relevant information about concepts and their meaning is distributed throughout an LDA. For example, procedural code, database schema, and/or parameter values can be extracted from data or obsolete documentation, and expertise of the system users or designers (where available) can be collected.

Davis and Aiken (2000) partition a major portion of the DRE literature into three areas: translation algorithms and methodologies, tools, and application-specific experiences. Translation algorithm development in early DRE efforts involved manual rearrangement or reformatting of data fields, which was inefficient and error-prone (Davis and Aiken 2000). Publication of the relational data model (Codd 1970) provided theoretical support for research in automated discovery of relational dependencies (Casanova and Sa 1983; Silva and Melkanoff 1981). In the early 1980s, the focus shifted to translation of relations to Entity-Relationship (E/R) diagrams (Dumpala and Arora 1981; Melkanoff and Zaniolo 1980). Given the early successes with translation using the relational data model, DRE translation techniques were applied to flat file databases (Casanova and Sa 1983; Davis and Arora 1985) within domains such as enterprise schemas (Klug 1980). The aforementioned problem of re-engineering legacy code to reveal data relationships and database schema was discussed by Nilsson in the context of COBOL code (Nilsson 1985).

DRE in the 1990s was enhanced by cross-fertilization with software engineering. Chikofsky (1990) developed a taxonomy for reverse engineering that includes DRE methodologies and also highlights available DRE tools. DRE formalisms were better defined, and the focus began to shift toward DRE's interaction with the human user (Hainaut 1991). The emphasis continued to be on the relational data model, in particular, extraction of E/R and schema from relational databases (Chiang et al. 1994; Markowitz and Makowsky 1990; Song and Froehlich 1995). Applications focus continued to be placed on legacy systems, including DoD applications (Aiken et al. 1994). Research in DRE tools proliferated, resulting in systems such as August-II (data model reverse engineering (Davis 1995)), DB-MAIN (programmable CASE tool for database engineering (Englebert and Hainaut 1999)), and tools for translation of relational databases (Chiang et al. 1994).

3.1.2 Program Analysis. An important trend in software engineering research is the use of program analysis or program comprehension. The original goal was to help programmers understand how existing programs work and how they would be affected by proposed modifications, but applications to data reverse engineering are straightforward. Several approaches have been proposed for program comprehension and, in the last few years, there has been a considerable effort to automate the same. The most important techniques include *(program) slicing* (Horwitz and Reps 1992), *cliché recognition* (Wills 1994), and *pattern matching* (Paul and Prakash 1994), besides the more conventional approaches of lexical and syntactic analysis. Slicing is a data flow analysis derivative that helps programmers understand what an existing program does and how it works by reducing the available code to only those lines that manipulate a certain set of program points of interest (e.g., input and output variables and their dependents).

Clichés are commonly used computational structures in programs. Examples of clichés include: list enumeration, binary search, and common data structures, such as hash table and priority queues. Since clichés have well-known properties and behavior, cliché recognition allows an experienced programmer to reconstruct the program's design and comprehend it. Commonly occurring programming patterns are encoded in terms of data and control flow constraints and stored in a cliché library. Recognition is achieved by parsing the flow graph of the program in accordance to the language grammar. Attribute and constraint checking with the cliché library is interleaved with the parsing.

Pattern matching is a technique that identifies interesting patterns and their dependencies in the code. For example, conditional control structures such as if..then..else, or case statements may encode business knowledge, whereas data type declarations and class or structure definitions can provide valuable information about the names, data types, and structure of concepts represented in an underlying database. Interesting programming patterns are stored in templates using a so-called 'pattern language'. Pattern matching works by transforming both the source code and the pattern templates into syntax trees. A code pattern recognizer performs the matching of patterns from the templates against patterns in the source code. Coupled with *program dependence graphs,* a language independent program representation, slicing, cliché recognition, and pattern matching are valuable tools for extracting semantic information from application code.

In the late 1990s, object-oriented DRE was explored in terms of discovering objects in legacy systems using function-, data-, and object-driven objectification (Wiggerts et al. 1997). Applications of DRE continued to grow, particularly in identification and remediation of the Y2K bug. The recent focus of DRE is more application-oriented, for example, mining of large data repositories (Dayani-Fard and Jurisica 1998), analysis of legacy systems (Hensley and Davis 2000) or network databases (Moh 2000), and extraction of business rules hidden within legacy systems (Shao and Pound 1999). Current research in the areas of software engineering and database systems focuses on developing powerful DRE tools, refining heuristics to yield fewer missing constructs, and developing techniques for reengineering legacy systems into distributed applications.

3.1.3 Data Mining.

Loosely speaking, the term data mining refers to the process of semi-automatically analyzing large data sets to find useful and interesting patterns. Like machine learning (Mitchell 1997) or statistical analysis (DeGroot and Schervish 2002), data mining attempts to extract previously unknown rules and patterns from the data. However, data mining also differs from machine learning and statistics in that it deals with large volumes of data stored primarily on magnetic disk. Hence, data mining can be used for knowledge extraction from databases.

As in the case of data reverse engineering and program analysis, there are important manual components in data mining, consisting of (1) preprocessing data to a form acceptable to the mining algorithms, and (2) post-processing of discovered patterns to find novel ones that could be useful. There may also be more than one type of pattern that can be extracted from a given database, and manual interaction may be needed to select useful types of patterns for further examination.

In general, data mining tasks can be classified into two categories: *descriptive* and *predictive.* Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. Examples of predictive data mining algorithms are classification algorithms such as decision tree classifiers, Bayesian classifiers, and classification based on regression (see Weiss and Kulikowski 1991) for a comprehensive introduction to classification and prediction methods). An example of descriptive mining is *association rule mining* (Agrawal et al. 1996), which is the extraction of rules showing attribute-value conditions that occur frequently together in a set of data. *Cluster analysis* (Jain et al. 1999) is another example of a descriptive data mining algorithm, which attempts to identify clusters of related points in the given data set. Unlike classification, which assumes knowledge about the classes, clustering analyzes data objects without knowledge of any class labels.

Instead, clustering is used to generate such labels. For a textbook-style coverage of data mining, including detailed descriptions of the above-mentioned algorithms, the reader is invited to refer to Han and Kamber (2001).

We believe that data mining, especially descriptive data mining, plays an important role in the extraction of knowledge that is stored as data in the information repository of an enterprise. Since this knowledge is based on current data values, which are likely to change over time, we call the extracted knowledge *situational knowledge* (SK) to distinguish it from static knowledge such as business rules. There are some important points to consider when attempting to extract situational knowledge:

- Extraction of situational knowledge requires a concise, abstract description of the data in the database.
- The most applicable mining algorithm is called concept description mining and produces a *generalization hierarchy* describing the database at different levels of granularity and from different angles.
- Given the fact that situational knowledge mining is based on the actual data rather than the metadata constraining the data, the results should be used for validation purposes (e.g., to validate the business rules extracted from application code) rather than as facts in themselves.
- Like other data mining processes, extraction of situational knowledge has to be monitored by a domain expert.

Various techniques have been developed to produce generalization hierarchies (e.g., Cai et al. 1991; Han et al. 1998). What is common to all of these techniques is that they abstract a large set of task-relevant data in a database from low conceptual levels to higher ones.

3.2 Knowledge composition: a mediation approach

The ability to extract knowledge using techniques such as DRE and data mining is not, by itself, a complete solution to the problems of providing useful information for supply chain coordination. Beyond an ability to connect to and provide basic translation of data in legacy systems in the supply chain, raw data must often be transformed to a form suitable for decision-making. Consider that much of the data used for operations in firms is detailed in nature, often mimicking accounting details or a detailed work breakdown structure. This data is too detailed for most supply chain decision support models (see for example, the models in Tayur et al. 1999). More broadly, we must compose the data needed as input for analysis tools from data used by applications with other purposes in mind. This knowledge composition may involve considerable processing of knowledge extracted from firms' legacy systems before it is usable by supply chain analysis tools. The need to aggregate detailed schedule data from a subcontractor for a master schedule is one example of such processing of extracted knowledge. Summary costs from detailed accounting data is another example.

There are several approaches to knowledge composition, ranging from well-known data transformations to machine learning and other artificial intelligence-based algorithms. While there are a variety of technical approaches, we have found the *mediation* approach a useful way of understanding and addressing the challenges of knowledge composition. Mediation is a value-added service (Wiederhold 1992) that acts as a layer between legacy sources and queries from users or applications. The services of mediators broadly include: Assembly or fusion of information from distributed data sources; transformation of that data into another form (often increasing the information density by combining or summarizing raw data); and cleansing the data to improve its reliability or accuracy (e.g., resolving conflicting information, identifying and interpolating missing data, providing a confidence level of the mediated information).

Each of the services of mediators is necessary when addressing the challenges of knowledge composition to support supply chain coordination. Knowledge about resource availability may reside in multiple sources within a single firm, requiring considerable data fusion or assembly to provide an answer to a (seemingly) simple query from a hub such as "do you have resources available on this date?" Similarly, the raw data of the firm may require translation from one syntax to another; a query about work crews may require considerable syntactic translation between firms' terminology and the syntax required by the querying application. Cleansing data is another large problem; one firm's detailed information system may support complex queries, but duplication of data across that firm's systems could be conflicting. Other firms in the supply chain may have less sophisticated information systems that cannot provide answers to queries or that require some interpolation of missing data to provide estimates.

Numerous integration projects have addressed aspects of the mediation approach. Multibase (Smith et al. 1981) and descendants, Mermaid (Templeton and al. 1987), Pegasus (Ahmed et al. 1993), HERMES (Palomaeki et al. 1993), SIMS (Arens et al. 1993), TSIMMIS (Chawathe et al. 1994; Hammer et al. 1995), Information Manifold (Kirk et al. 1995; Levy et al. 1996), and, more recently, XPERANTO (Shanmugasundaram et al. 2001 and Xyleme 2001) are representatives that exemplify the current state-of-the-art. Despite all of their substantial contributions to the field of data integration and mediation, none of these projects have dealt with integrating complex data from semantically rich sources. Extending the ability of mediators to process heterogeneous information with some intelligence remains a fundamental challenge.

The mediation approach and associated challenges can be seen via a generic mediation architecture (depicted in Figure 9.3) designed to support complex queries for supply chain coordination. The mediator consists of a Query Processing Module (represented by the dashed box on the left-hand side of the mediator) responsible for decomposing a request for data into the subqueries which are submitted to the relevant data sources, and a Data Merge Engine (represented by the dashed box on the right-hand side of the mediator) responsible for integrating, cleansing, and reconciling the result data that is returned from the legacy sources.

Within the Query Processing module, the *Query Translation component* parses and translates the incoming queries (which may be formulated in a high-level language such as SQL) into the internal format of the mediator. The *Query Decomposition component* performs the query rewriting (e.g., Genesereth and Duschka 1997) of the client query into one or more firm-specific mediated subqueries. This re-writing is necessary since the data model and representation used by the firms is different from the integrated view of the data that is supported by the mediator. As a result, the mediated subqueries are expressed using the schema and terminology used by the underlying firms.

When the result is returned from the sources (typically via wrappers), the Data Merge Engine produces the integrated result that is returned to the client application. The *Result Fusion* component joins related data based on a set of merging rules. Data restructuring is done by the *Cleansing & Reconciliation* component and includes the removal of duplicate information, resolution of conflicts, as well as the aggregation and grouping of information into high-level units.

The ability to decompose queries and turn them into subqueries specific to the multiple legacy sources, together with the ability to fuse multiple results into an answer to the original query, requires considerable sophistication on the part of mediator. A basic issue that must be addressed in the construction of mediators is the specification of the internal language of the mediator. The internal syntax must be expressible enough to cover the projected range of queries from applications and users. With regard to supply chain coordination problems, we believe

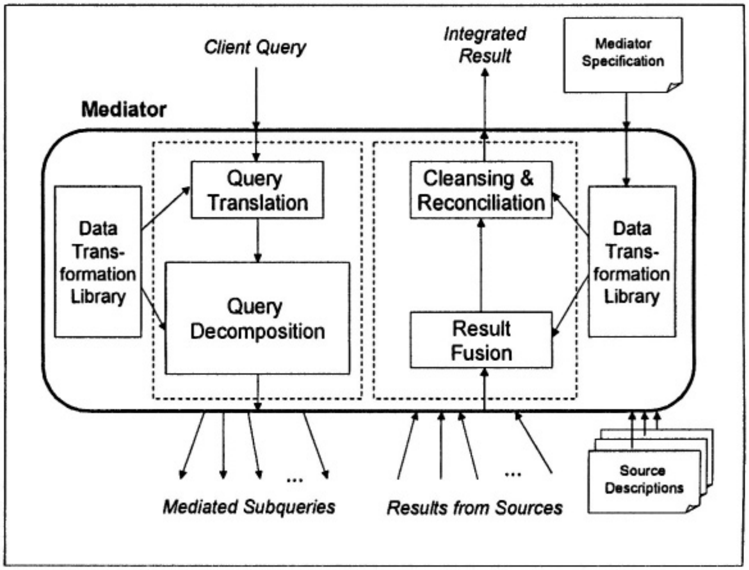


Figure 9.3. Mediation architecture for complex querying (e.g., supply chain applications).

that traditional data models (e.g., relational, object-oriented) are too rigid and structured to be able to represent knowledge efficiently. Process type knowledge is inherently dynamic in nature. Semistructured data (Hammer et al. 1997b; Suciu 1998) has emerged as a means for representing and exchanging heterogeneous data. Since semistructured data models represent data as a graph, modeling of a wide variety of data formats is simplified. The World Wide Web Consortium (W3C) has defined a standard data model and exchange syntax for semistructured data based on XML (Connolly 1997) called the Resource Description Framework (RDF) (Lassila and Swick 1999). Efforts are under way to use RDF to express semantically rich data; most notably DARPA's Agent Markup Language (DAML) program (DARPA 2000), for defining ontology languages in RDF. The advantage of using RDF is the ability to leverage emerging storage, query, and integration facilities for XML. In addition, the combination of RDF and DAML provides a very general approach to storing and manipulating processes since they can be viewed, stored, queried and exchanged as XML data using available tools. As such, we promote the use of RDF-based frameworks for the internal language of supply-chain/process type mediators.

4. Implementation Example: SEEK - Scalable Extraction of Enterprise Knowledge

Much of the research to date regarding knowledge extraction and composition has focused on single approaches and/or labor-intensive implementations that require considerable time and expertise to implement. Current techniques do not scale to the problem of extracting and composing knowledge from legacy sources to support supply chain coordination. There is a need for an integrated approach to enable rapid, seamless extraction and composition of source data and knowledge from large numbers of physically and semantically heterogeneous sources. We have been developing such an integrated approach known as SEEK - Scalable Extraction of Enterprise Knowledge - that significantly extends current techniques. Our novel approach uses encoded domain knowledge to allow rapid connection to and discovery of data and knowledge across a large number of heterogeneous sources. Due to our use of domain knowledge, it is important to note that SEEK is not a general-purpose toolkit. Rather, it allows scalable extraction of specific forms of knowledge.

In this section we provide a high level overview of the SEEK architecture and approach in Section 4.1. A more detailed discussion of the current implementation and algorithms are presented in Sections 4.2 and 4.3. Elements of Section 4 are drawn from O'Brien and Hammer (2001) and O'Brien et al. (2002). Readers may wish to refer to those articles for further discussion, particularly with regard to supply chains in project environments.

4.1 SEEK architecture and approach

A high-level view of the core SEEK architecture is shown in Figure 9.4. SEEK follows established mediation/wrapper methodologies (e.g., TSIMMIS (Chawathe et al. 1994), InfoSleuth (Bayardo et al. 1996)) and provides a software middleware layer that bridges the gap between legacy information sources and decision makers/support tools.

SEEK works as follows. During run-time, the *analysis module* processes queries from the end-users (e.g. decision support tools and analysts) and performs knowledge composition including basic mediation tasks and post-processing of the extracted data. Data communication between the analysis module and the legacy sources is provided by the *wrapper component*. The wrapper translates SEEK queries into access commands understood by the source and converts native source results into SEEK's unifying internal language (e.g., XML/RDF).

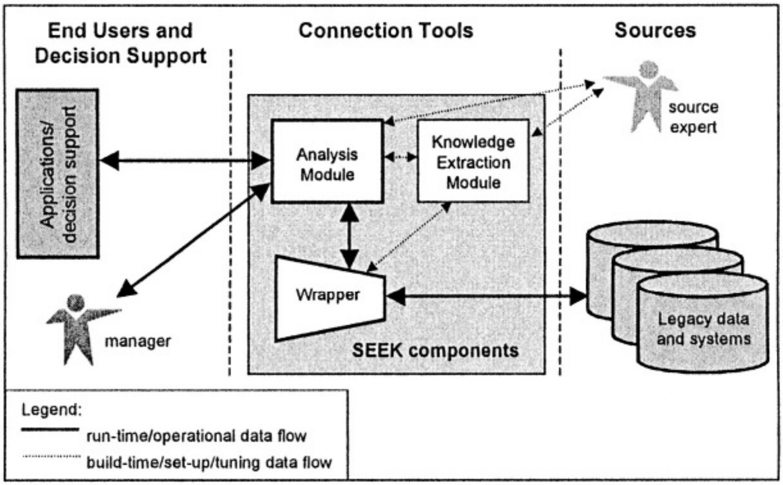

End Users and Decision Support
Connection Tools
Sources
Applications/ decision support
Analysis Module
Knowledge Extraction Module
source expert
Wrapper
Legacy data and systems
manager
SEEK components
Legend:
run-time/operational data flow
build-time/set-up/tuning data flow

Figure 9.4. High-level SEEK architecture and relation to legacy sources and querying applications.

Prior to answering queries, SEEK must be configured. This is accomplished semi-automatically by the *knowledge extraction module* that directs wrapper and analysis module configuration during build-time. The wrapper must be configured with information regarding communication protocols between SEEK and legacy sources, access mechanisms, and underlying source schemas. The analysis module must be configured with information about source capabilities, available knowledge and its representation. We are using a wrapper generation toolkit (Hammer et al. 1997a) for fast, scalable, and efficient implementation of customized wrappers. To produce a SEEK-specific representation of the operational knowledge in the sources, we are using domain specific templates to describe the semantics of commonly used structures and schemas. Wrapper configuration is assisted by a source expert to extend the capabilities of the initial, automatic configuration directed by the templates in the knowledge extraction module. Use of domain experts in template configuration is particularly necessary for poorly formed database specifications often found in older legacy systems. Furthermore, the knowledge extraction module also enables step-wise refinement of templates and wrapper configuration to improve extraction capabilities.

It is important to comment on the ability of the SEEK architecture to enhance security of legacy data. The SEEK architecture allows secure, privacy constrained filtering of legacy data in two ways. First, as an intermediary between a hub (e.g., the Application/Decision Support module in Figure 9.4) and legacy source, it restricts access to firm data and answers only specific queries. For example, the analysis module may respond to a query about resource availability on given dates with a simple affirmative or negative answer. The hub (and whoever operates the hub, such as an assembler sourcing parts from a supplier or a general contractor requesting information from a subcontractor) need never see the firm's raw data concerning resource allocation to other projects. A second aspect of security is that the firm can limit access to the legacy data available to the SEEK wrapper component. An access layer controls the overall permissions and availability of data. It is likely that the access layer will not be SEEK specific but be a general directory services platform that controls both internal and external access to a firm's information systems. As firms are justifiably reluctant to share details concerning operations, costs, etc., the existence of an intermediary between their legacy information and requester is both an important source of security and is an enabler for information integration in practice.

4.2 Current implementation overview

Our discussion above provides a high level overview of the SEEK architecture. We believe that the modular structure of the architecture provides a generalized approach to knowledge extraction that is applicable in many circumstances. That said, it is useful to provide more details of the current implementation architecture before delving into the details of data extraction algorithms. As such, a more detailed schematic is shown in Figure 9.5. SEEK applies Data Reverse Engineering (DRE) and Schema Matching (SM) processes to legacy database(s) to produce a source wrapper for a legacy source. The source wrapper will be used by another component (not shown in Figure 9.5) wishing to communicate and exchange information with the legacy system. We assume that the legacy source uses a database management system for storing and managing its enterprise data.

First, SEEK generates a detailed description of the legacy source, including entities, relationships, application-specific meanings of the entities and relationships, business rules, data formatting and reporting constraints, etc. We collectively refer to this information as *enterprise knowledge.* The extracted enterprise knowledge forms a knowledgebase that serves as input for subsequent steps outlined below. In order to extract this enterprise knowledge, the DRE module shown on the left of Figure 9.5 connects to the underlying DBMS to extract schema information (most data sources support at least some form of Call-Level

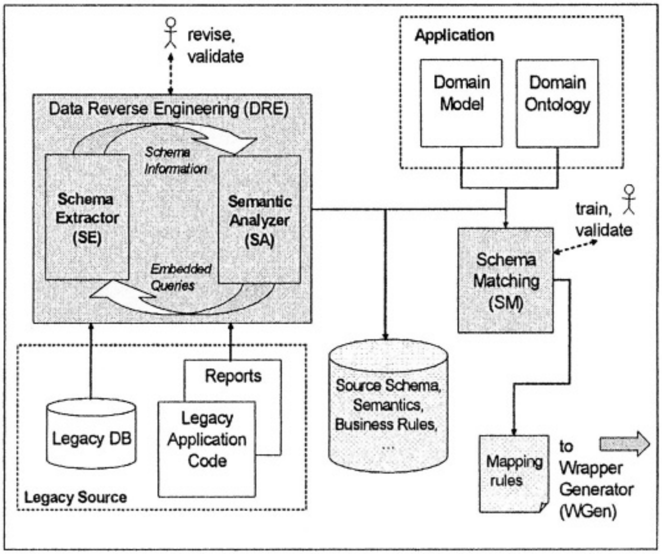


Figure 9.5. Schematic diagram of the conceptual architecture of SEEK's knowledge extraction algorithm.

Interface such as JDBC). The schema information from the database is semantically enhanced using clues extracted by the semantic analyzer from available application code, business reports, and, in the future, perhaps other electronically available information that may encode business data such as e-mail correspondence, corporate memos, etc.

Second, the semantically enhanced legacy source schema must be mapped into the domain model (DM) used by the application(s) that want(s) to access the legacy source. This is done using a schema matching process that produces the mapping rules between the legacy source schema and the application domain model. In addition to the domain model, the schema matching module also needs access to the domain ontology (DO) describing the model.

Third, the extracted legacy schema and the mapping rules provide the input to the wrapper generator (not shown in Figure 9.5), which produces the source wrapper.

The three preceding steps can be formalized as follows. At a high level, let a legacy source L be denoted by the tuple $L = (DB_L, S_L, D_L, Q_L)$, where DB_L denotes the legacy database, S_L denotes its schema, D_L the data and Q_L a set of queries that can be answered by DB_L. Note that the legacy database need not be a relational database, but can include text, flat file databases, and hierarchically formatted information. S_L is expressed by the data model DM_L.

We also define an application via the tuple $A = (S_A, Q_A, D_A)$, where S_A denotes the schema used by the application and Q_A denotes a collection of queries written against that schema. The symbol D_A denotes data that is expressed in the context of the application. We assume that the application schema is described by a domain model and its corresponding ontology (as shown in Figure 9.5). For simplicity, we further assume that the application query format is specific to a given application domain but invariant across legacy sources for that domain.

Let a *legacy source wrapper W* be comprised of a query transformation

$$f\mathrm{w}^{\mathrm{Q}} : \mathrm{Q_A} \mapsto \mathrm{Q_L} \tag{9.1}$$

and a data transformation

$$f\mathrm{w}^{\mathrm{D}} : \mathrm{D_L} \mapsto \mathrm{D_A} \tag{9.2}$$

where the *Q*'s and *D*'s are constrained by the corresponding schemas.

The *SEEK knowledge extraction process* shown in Figure 9.5 can now be stated as follows. We are given S_A and Q_A for an application wishing to access legacy database DB_L whose schema S_L is unknown. Assuming that we have access to the legacy database DB_L as well as to application code C_L accessing DB_L, we first infer S_L by analyzing DB_L and C_L, then use S_L to infer a set of mapping rules *M* between S_L and S_A, which are used by a wrapper generator *WGen* to produce $(f\mathrm{w}^{\mathrm{Q}}, \mathrm{fw}^{\mathrm{D}})$. In short:

$$\begin{aligned} DRE : (DB_L, C_L) &\mapsto S_L && (9.3)\\ SM : (S_L, S_A) &\mapsto M && (9.4)\\ WGen : (Q_A, M) &\mapsto (f\mathrm{w}^{\mathrm{Q}}, \mathrm{fw}^{\mathrm{D}}) && (9.5) \end{aligned}$$

In the next section, we provide a description of our *DRE* algorithm (encoded by Equation (9.3)) that deploys schema extraction and semantic analysis. Our work on *SM* is still in the early stages of development and will be presented in subsequent publications. An excellent summary of the current state-of-the-art in schema matching can be found in Rahm and Bernstein (2001). A description of a specific instantiation of *WGen,* which goes beyond the scope of this article, can be found in Hammer et al. (1997a).

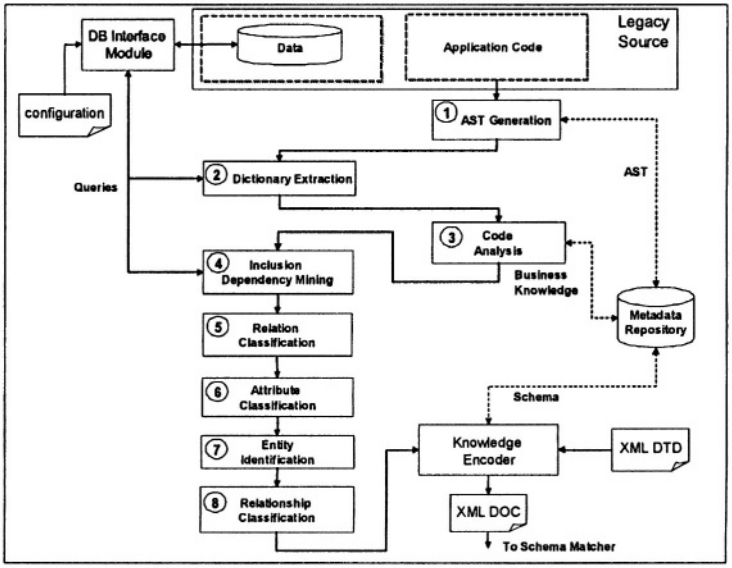


Figure 9.6. Conceptual overview of the DRE algorithm.

4.3 Example: SEEK DRE implementation for extracting resource and schedule data

As discussed in Section 3, data reverse engineering (DRE) is defined as the application of analytical techniques to one or more legacy data sources to elicit structural information (e.g., term definitions, schema definitions) from the legacy source(s) in order to improve the database design or produce missing schema documentation. Thus far in SEEK, we are applying DRE to relational databases only. However, since the relational model has only limited semantic expressability, in addition to the schema, our DRE algorithm generates an E/R-like representation of the entities and relationships that are not explicitly defined in the legacy schema (but which exist implicitly). Our approach to data reverse engineering for relational sources is based on existing algorithms by Chiang (1995) and Chiang et al. (1994) and Petit et al. (1996). However, we have improved their methodologies in several ways, most importantly to reduce the dependency on human input and to eliminate some of the limitations of their algorithms (e.g., consistent naming of key attributes, legacy schema in 3-NF). We highlight our improvements during our description of the DRE algorithm below.

Our DRE algorithm is divided into schema extraction and semantic analysis, which operate in an interleaved fashion. An overview of the two algorithms, which are comprised of eight steps, is shown in Figure 9.6. In addition to the modules that execute each of the eight steps, the architecture in Figure 9.6 includes three support components:

1 The configurable *Database Interface Module* (upper-right hand corner), which provides connectivity to the underlying legacy source. Note that this component is the ONLY source-specific component in the architecture: in order to perform knowledge extraction from different sources, only the interface module needs to be modified to be compatible with the source.

2 The *Knowledge Encoder* (lower right-hand corner) represents the extracted knowledge in the form of an XML document, which can be shared with other components in the SEEK architecture (e.g., the semantic matcher).

3 The *Metadata Repository* is internal to DRE and is used to store intermediate, run-time information needed by the algorithms, including user input parameters, the abstract syntax tree for the code (e.g., from a previous invocation), etc.

The DRE extraction algorithm uses schema information stored in the legacy database to extract information about how the data is organized and stored. If available, DRE also analyzes the application code that uses the database to detect clues about the meaning and usage of the database structures. For example, output statements in the application code usually have detailed comments about the data that is to be printed or received; these semantically rich comments can be associated with the corresponding structures in the database and allow DRE to enhance the structural information with semantics. This semantically enhanced structural information will later be used by *SM* and *WGen* to build a functional source wrapper.

The next section describes the highlights of each of the eight DRE steps. Readers wishing to omit the technical details can skip this section and go to Section 4.3.2 for a summary of the outcome of DRE.

4.3.1 The 8-step DRE algorithm. We briefly highlight each of the eight steps and related activities outlined in Figure 9.6 using an example from our construction supply chain testbed. For a detailed description of our algorithm, refer to Hammer et al. (2002). For simplicity, we assume without loss of generality or specificity that only the following relations exist in the MS-Project application,[1] which will be discovered using DRE. Due to space constraints, we are also limiting our description of the source schema to the important key attributes (for a description of the entire schema refer to Microsoft Corp. 2000). Relation names are in bold font, primary key attributes are underlined:

MSP-Project (PROJ ID, ...)
MSP-Availability(PROJ ID, TASK UID, ...)
MSP-Resources (PROJ ID, RES UID, ...)
MSP-Tasks (PROJ ID, TASK UID, ...)
MSP-Assignment (PROJ ID, ASSN UID, ...)

In order to illustrate the code analysis and how it enhances the schema extraction, we refer the reader to the following C code fragment representing a simple, hypothetical interaction with the MS Project database.

```
char *aValue, *cValue;
int flag = 0;
int bValue = 0;
EXEC SQL SELECT A,C INTO :aValue, :cValue
FROM Z WHERE B =: bValue;
if (cValue < aValue) { flag = 1; }
printf("Task Start Date %s ", aValue);
printf("Task Finish Date %s ", cValue);
```

Step 1: AST Generation
We start by creating an Abstract Syntax Tree (AST) shown in Figure 9.7. The AST will be used by the semantic analyzer for code exploration during Step 3. Our objective in AST generation is to be able to associate "meaning" with program variables. Format strings in input/output statements contain semantic information that can be associated with the variables in the input/output statement. A program variable in turn may be associated with a column of a table in the underlying legacy database.

Step 2. Dictionary Extraction
The goal of Step 2 is to obtain the *relation* and *attribute names* from the legacy source. This is done by querying the data dictionary, stored in the underlying database in the form of one or more system tables.

[1] It is useful to note that the use of MS-Project as a test example represents a significant challenge because of the complexity of the underlying database schema, which is used by MS Project as a repository and which serves as the starting point for our DRE process.

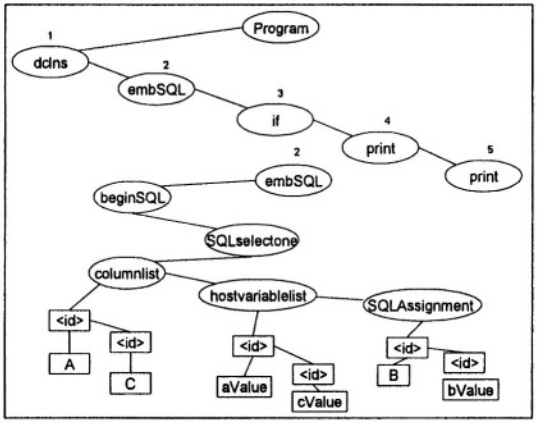


Figure 9.7. Application-specific code analysis via AST decomposition and code slicing. The direction of slicing is backward (forward) if the variable in question is in an output (respectively input or declaration) statement.

Otherwise, if primary key information cannot be retrieved directly from the data dictionary, the algorithm passes the set of candidate keys along with predefined "rule-out" patterns to the code analyzer. The code analyzer searches for these patterns in the application code and eliminates those attributes from the candidate set, which occur in these "rule-out" patterns. The rule-out patterns, which are expressed as SQL queries, occur in the application code whenever the programmer expects to select a SET of tuples. If, after the code analysis, not all primary keys can be identified, the reduced set of candidate keys is presented to the user for final primary key selection.

Result. In the example DRE application, the following relations and their attributes were obtained from the MS-Project database:

MSP-Project (PROJ ID, ...)
MSP-Availability(PROJ ID, AVAIL UID, ...)
MSP-Resources (PROJ ID, RES UID, ...)
MSP-Tasks (PROJ ID, TASK UID, ...)
MSP-Assignment (PROJ ID, ASSN UID, ...)

Step 3: Code Analysis

The objective of Step 3, code analysis, is twofold: (1) augment entities extracted in Step 2 with domain semantics, and (2) identify business rules and constraints not explicitly stored in the database, but which may be important to the wrapper developer or application program accessing the legacy source. Our approach to code analysis is based on code slicing (Horwitz and Reps 1992) and pattern matching (Paul and Prakash 1994).

The first step is the *pre-slicing* step. From the AST of the application code, the pre-slicer identifies all the nodes corresponding to input, output and embedded SQL statements. It appends the statement node name, and identifier list to an array as the AST is traversed in pre-order. For example, for the AST in Figure 9.7, the array contains the following information depicted in Table 9.1. The identifiers that occur in this data structure maintained by the pre-slicer form the set of slicing variables.

Node Number	Statement	Text String (for print nodes)	Identifiers	Direction of Slicing
2	embSQL (Embedded SQL node)	——	aValue cValue	Backward

Table 9.1. Information maintained by the pre-slicer.

The code slicer and analyzer, which represent Steps 2 and 3 respectively, are executed once for each slicing variable identified by the pre-slicer. In the above example, the slicing variables that occur in SQL and output statements are aValue and cValue. The direction of slicing is fixed as backward or forward depending on whether the variable in question is part of an output (backward) or input (forward) statement. The slicing criterion is the exact statement (SQL or input or output) node that corresponds to the slicing variable.

During the code slicing sub-step we traverse the AST for the source code and retain only those nodes that have an occurrence of the slicing variable in the sub-tree. This results in a reduced AST, which is shown in Figure 9.8.

During the analysis sub-step, our algorithm extracts the information shown in Table 9.2, while traversing the reduced AST in pre-order.

1 If a *dcln* node is encountered, the data type of the identifier can be learned.

2 *embSQL* contains the mapping information of identifier name to corresponding column name and table name in the database.

3 *Printf/scanf* nodes contain the mapping information from the text string to the identifier. In other words we can extract the 'meaning' of the identifier from the text string.

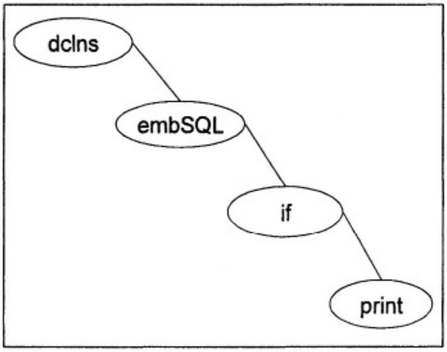


Figure 9.8. Reduced AST.

Identifier Name	Meaning	Possible Business Rule	Data Type	Column Name in Source	Table Name in Source

aValue	Task Start	if(cValue < aValue)
		{
		}
			Char* ≥ string	A	Z

cValue	Task Finish	if(cValue < aValue)
		{
		}
			Char* ≥ string	C	Z

Table 9.2. Information maintained by the pre-slicer.

The results of the analysis sub-step are appended to a result report file. After the code slicer and analyzer have been invoked on every slicing variable identified by the pre-slicer, the results report file is presented to the user. The user can base his/her decision of whether to perform further analysis based on the information extracted so far. If the user decides not to perform further analysis, code analysis passes control to the inclusion dependency detection module.

It is important to note that we identify enterprise knowledge by matching templates against code fragments in the AST. So far, we have developed patterns for discovering business rules which are encoded in loop structures, conditional statements, or mathematical formulae, which in turn are encoded in loop structures or assignment statements. Note that the occurrence of an assignment statement itself does not necessarily indicate the presence of a mathematical formula, but the likelihood increases significantly if the statement contains one of the slicing variables.

Step 4. Discovering Inclusion Dependencies

Following extraction of the relational schema in Step 2, the goal of Step 4 is to identify constraints to help classify the extracted relations, which represent both the real-world entities and the relationships among them. This is achieved using inclusion dependencies, which indicate the existence of inter-relational constraints including class/subclass relationships.

Let A and B be two relations, and X and Y be attributes or a set of attributes of A and B respectively. An inclusion dependency A.X << B.Y denotes that a set of values appearing in A.X is a subset of B.Y. Inclusion dependencies are discovered by examining all possible subset relationships between any two relations A and B in the legacy source.

Without additional input from the domain expert, inclusion dependencies can be identified in an exhaustive manner as follows: for each pair of relations A and B in the legacy source schema, compare the values for each non-key attribute combination X in B with the values of each candidate key attribute combination Y in A (note that X and Y may be single attributes). An inclusion dependency B.X << A.Y may be present if:

1 X and Y have same number of attributes.

2 X and Y must have pair-wise domain compatibility.

3 B.X $\subseteq$ A.Y

In order to check the subset criterion (3), we have designed the following generalized SQL query templates, which are instantiated for each pair of relations and attribute combinations and run against the legacy source:

```
C1 =                        C2 =
SELECT count (*)            SELECT count (*)
FROM R1                     FROM R2
WHERE U NOT IN              WHERE V NOT IN
   (SELECT V                   (SELECT U
   FROM R2);                   FROM R1);
```

If C1 is zero, we can deduce that there may exist an inclusion dependency R1.U << R2.V; likewise, if C2 is zero there may exist an inclusion dependency R2.V << R1.U. Note that it is possible for both C1 and C2 to be zero. In that case, we can conclude that the two sets of attributes U and V are equal.

The worst-case complexity of this exhaustive search, given N tables and M attributes per table (NM total attributes), is $O(N^2M^2)$.

However, we reduce the search space in those cases where we can identify equi-join queries in the application code (during semantic analysis). Each equi-join query allows us to deduce the existence of one or more inclusion dependencies in the underlying schema. In addition, using the results of the corresponding count queries we can also determine the directionality of the dependencies. This allows us to limit our exhaustive searching to only those relations not mentioned in the extracted queries.

Result: Inclusion dependencies are as follows:
1 MSP-Assignment(Task_uid,Proj_ID) << MSP-Tasks (Task_uid,Proj_ID)
2 MSP-Assignment(Res_uid.Proj_ID) << MSP-Resources(Res_uid,Proj_ID)
3 MSP-Availability (Res_uid,Proj_ID) << MSP-Resources (Res_uid,Proj_ID)
4 MSP-Resources (Proj_ID) << MSP-Project (Proj_ID)
5 MSP-Tasks (Proj_ID) << MSP-Project (Proj_ID)
6 MSP-Assignment (Proj_ID) << MSP-Project (Proj_ID)
7 MSP-Availability (Proj_ID) << MSP-Project (Proj_ID)

The last two inclusion dependencies are removed since they are implicitly contained in the inclusion dependencies listed in lines 2, 3 and 4 using the transitivity relationship.

Step 5. Classification of the Relations

When reverse-engineering a relational schema, it is important to understand that due to the limited expressability of the relational model, all real-world entities are represented as relations irrespective of their types and role in the model. The goal of this step is to identify the different "types" of relations, some of which correspond to actual real-world entities while others represent relationships among them.

In this step, all the relations in the database are classified into one of four types - *strong, regular, weak,* or *specific.* Identifying different relations is done using the primary key information obtained in Step 2 and the inclusion dependencies from Step 4. Intuitively, a strong entity-relation represents a real-world entity whose members can be identified exclusively through its own properties. A weak entity-relation represents an entity that has no properties of its own which can be used to identify its members. In the relational model, the primary keys of weak entity-relations usually contain primary key attributes from other (strong) entity-relations. Both regular and specific relations are relations that represent relationships between two entities in the real world (rather then the entities themselves). However, there are instances when not all of the entities participating in an (n-ary) relationship are present in the database schema (e.g., one or more of the relations were deleted as part of the normal database schema evolution process). While reverse engineering the database, we identify such relationships as special relations.

Result:
Strong Entities: MSP-Projects
Weak Entities: MSP-Resources, MSP-Tasks, MSP-Availability
Regular Relationship: MSP-Assignment

Step 6. Classification of the Attributes
We classify attributes as (a) PK or FK (from DRE-1 or DRE-2), (b) Dangling (DKA) or General (GKA) Key, or (c) Non-Key (NKA). For clarification, a set of one or more attributes classified as PK (primary key) attributes represent a unique identifier for the relation to which they belong. Attributes classified as FK (foreign key) attributes represent "copies" of the primary keys from other relations and express a relationship between related tuples in two relations (example of an inter-relational constraint). Dangling attributes are those attributes that also appear in other relations but are not part of the primary key. General attributes, also known as candidate key attributes, are attributes that can also uniquely identify tuples in a relation but have not been designated as primary key (e.g., both ss# and the combination of birthdate and name could be used to uniquely identify persons; however, only one can be designated as PK). All other attributes are termed non-key attributes.

Result: Table 9.3 illustrates attributes obtained from the example legacy source.

	PKA	DKA	GKA	FKA	NKA
MS-Project	Proj_ID				All Remaining Attributes
MS-Res.	Proj_ID	Res_uid			
MS-Tasks	Proj_ID	Task_uid			
MS-Avail.	Proj_ID	Avail_uid		Res_uid + Proj_ID	
MS-Assign.	Proj_ID		Assn_uid	Res_uid + Proj_ID, Task_uid + Proj_ID	

Table 9.3. Example of attribute classification from MS-Project legacy source.

Step 7. Identify Entity Types
Strong (weak) entity relations obtained from Step 5 are directly converted into strong (weak) entities.

Result: The following entities were classified:

Strong Entities:

MSP-Project with Proj_ID as its key.

Weak entities:

MSP-Tasks with Task_uid as key and MSP-Project as its owner.
MSP-Resources with Res_uid as key and MSP-Project as its owner.
MSP-Availability with Avail_uid as key and MSP-Resources as owner.

Step 8. Identify Relationship Types

The inclusion dependencies discovered in Step 4 form the basis for determining the relationship types among the entities identified above. This is a two-step process:

1 Identify relationships present as relations in the relational database. The relation types (regular and specific) obtained from the classification of relations (Step 5) are converted into relationships. The participating entity types are derived from the inclusion dependencies. For completeness of the extracted schema, we may decide to create a new entity when conceptualizing a specific relation. The cardinality between the entities is M:N.

2 Identify relationships among the entity types (strong and weak) that were not present as relations in the relational database, via the following classification.

- *IS-A relationships* can be identified using the PKAs of strong entity relations and the inclusion dependencies among PKAs. The cardinality of the IS-A relationship between the corresponding strong entities is 1:1.
- *Dependent relationship:* For each weak entity type, the owner is determined by examining the inclusion dependencies involving the corresponding weak entity-relation. The cardinality of the dependent relationship between the owner and the weak entity is 1:N.
- *Aggregate relationships:* If the foreign key in any of the regular and specific relations refers to the PKA of one of the strong entity relations, an aggregate relationship is identified. The cardinality is either 1:1 or 1:N.
- *Other binary relationships:* Other binary relationships are identified from the FKAs not used in identifying the above relationships. If the foreign key contains unique values, the cardinality is 1:1, otherwise the cardinality is 1:N.

Result:

We discovered 1:N binary relationships between the following weak entity types:

Between MSP-Project and MSP-Tasks
Between MSP-Project and MSP-Resources
Between MSP-Resources and MSP-Availabilty

Since two inclusion dependencies involving MSP-Assignment exist (i.e., between Task and Assignment and between Resource and Assignment), there is no need to define a new entity. Thus, MSP-Assignment becomes an M:N relationship between MSP-Tasks and MSP-Resources.

4.3.2 Outcome of DRE. At the end of Step 8, DRE has extracted the following schema information from the legacy database:

- Names and classification of all entities and attributes
- Primary and foreign keys.
- Data types (i.e., domain constraints).
- Simple constraints (e.g., unique) and explicit assertions.
- Relationships and their cardinalities.
- Enterprise knowledge encoded in rules and constraints (business rules).

A conceptual overview of the extracted schema is represented by the Entity-Relationship diagram shown in Figure 9.9 (business rules not shown). Note that we are using the E/R data model to represent the extracted information since it enables us to represent structure and semantics from legacy sources that are based on more expressive representations than what is possible in the relational model, such as legacy sources using object-based (e.g., persistent C++) or semistrutured data models (e.g., XML).

The diagram in Figure 9.9 is an accurate representation of the information encoded in the original MS Project relational schema. For example, in the schema in Figure 9.9, rectangles represent entities, which correspond to relations in the relational data model. Diamonds represent relationships between entities (or relations), which are also encoded as relations in the relational model. The alphanumeric values next to the lines representing the relationships indicate the cardinality constraints (1-1, 1-N, N-M). Ovals represent attributes. In keeping with the running example, we only show the key attributes (underlined) for each entity. Entities and relationships marked with double lines represent weak entities/relationships, which rely on the dependent strong entity for identification. For example, since tasks cannot exist without projects,

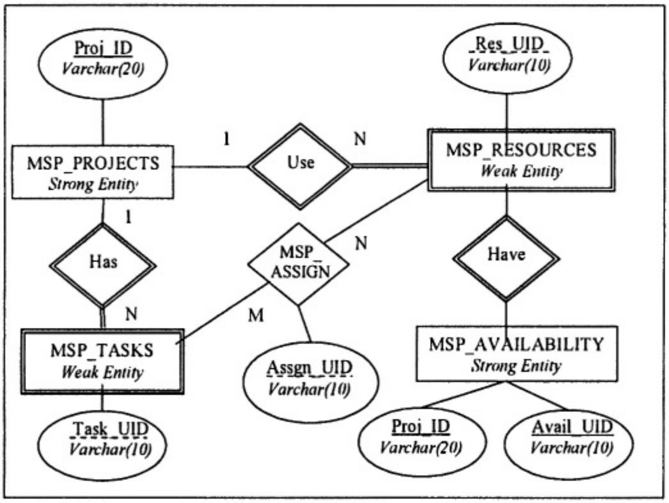


Figure 9.9. E/R diagram representing the extracted schema.

MSP_Tasks uses the Proj_ID key from the corresponding strong entity MSP_Projects to identify individual tasks (together with its own key, Task_UID). Attributes that are underlined with a dashed line represent partial keys for weak entities that need to be augmented with the key attributes from the corresponding strong entities. Finally, in our running example, some of the entities that represent relationships have attributes of their own (e.g., Assn_UID of relationship MSP-Assignment). However, in the relational representation, the corresponding (relationship) relations contain both key attributes from the relations that are being linked. For information on how to translate an E/R diagram into a relational schema, refer to Elmasri and Navathe (2000).

The figure also shows the data types for each attribute, which in this example, are strings of varying length (Varchar(N)). Not shown are unique and semantic constraints such as "the values of the attribute avail_units in MSP_Availability must always be greater than 0." Constraints can either be enforced by the underlying database management system or by the application code. DRE is capable of extracting both types of constraints.

Besides a description of the *structure* of the legacy database, DRE also infers the meaning of the attributes using clues about their usage and contents found in the application code per Step 3 described in the previous section. The meanings for the attributes shown in Figure 9.9 are listed in Table 9.4.

Attribute	Meaning	Corresponding Relation(s)
Proj_ID	Unique Project Identifier	MSP-Project, MSP-Resources, MSP-Tasks, MSP-Assignments, MSP-Availability
AVAIL_ID	Unique availability identifier	MSP-Availability
RES_UID	Unique resource identifier	MSP-Resources
TASK_UID	Unique task identifier	MSP-Tasks
ASSN_UID	Unique assignment identifier	MSP-Assignment

Table 9.4. Extracted meaning of selected attributes from the running example.

Using the 8-step algorithm described in Section 4.3.2, we are able to extract significant semantic knowledge encoded in the schema of a relational database. As described in Section 4.2, this knowledge is subsequently used by *SM* to produce mapping rules for translating data and queries between the legacy source and the domain model of the application needing to access the data in the legacy firm. The mapping rules in turn are used by *WGen* to produce a source-specific wrapper which is used by the analysis module to retrieve data from the legacy source.

Putting the previous discussion on DRE into perspective, it helps to remember that motivation for SEEK stems from the need for coordination of processes across multiple firms. However, the large number of firms on a project makes it likely that there will be a high degree of physical and semantic heterogeneity in their legacy systems, making it difficult to connect firms' data and systems with enterprise level decision support tools. It is the role of the SEEK system to (1) securely extract data and knowledge resident in physically and semantically heterogeneous legacy sources and (2) compose that knowledge to support queries not natively supported by the sources. The technologies outlined so far achieve the first part of the two goals stated above, namely they provide the ability to develop software for accessing and retrieving data from a wide variety of legacy sources. Future research efforts in the SEEK project are aimed at achieving the second goal, namely the secure composition of knowledge to help answer complex queries not natively supported by the legacy source/wrapper.

5. Future Research Directions

Despite the promising initial results of the SEEK knowledge extraction methodology, our work motivates several research areas to extend our ability to rapidly and accurately extract knowledge from legacy sources. Specifically, discovery and extraction of knowledge require an understanding of the underlying domain, something which is inherently subjective and for which current computer systems are not well suited, given the complexity of the reasoning processes and the large number of facts involved (Morey et al. 2000). As a result, most knowledge extraction systems rely on a human domain expert to at least supervise the extraction process by validating the extracted knowledge, making corrections if necessary, and/or providing training data for learning-based systems. This dependency naturally limits the scalability of the tools as well as the accuracy and quality of the knowledge that can be automatically extracted. Knowledge extraction systems that can automatically tune themselves to the specifics of underlying sources and extract high quality knowledge accurately and without human involvement remain a fundamental research challenge.

We envision three areas of research to extend the integrated approach to knowledge extraction and composition presented by SEEK: (1) Increasing intelligence about capabilities of legacy sources, either through extensions to SEEK or through Web-service like advertising of source capabilities. Such development will make SEEK and related approaches more robust and more scalable in terms of speed of deployment. (2) Bio-computing or evolutionary approaches to enhance the speed and quality of configuration. Such approaches may make it possible for near fully-automatic configuration. (3) Complementing the first two areas, a third approach is to integrate and extend research on knowledge representation at the user level, particularly with regard to visualization of the data. Such tools can augment the capabilities of domain experts (e.g., managers) in parsing data available from the SEEK system during run-time. Such extensions also increase the ability of managers to aid configuration during build-time, obviating much of the need for technical experts. The relationship between these extensions and the current work and approach is summarized in Figure 9.10.

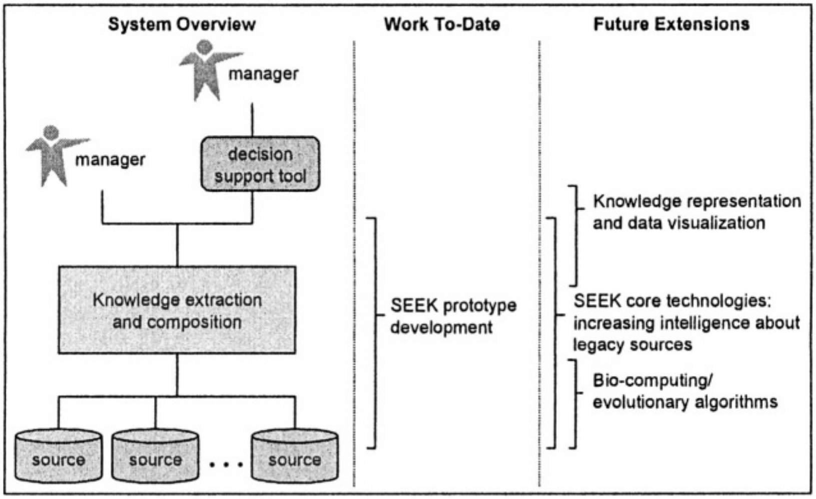

System Overview
Work To-Date
Future Extensions
manager
manager
decision support tool
Knowledge extraction and composition
source
source
...
source
SEEK prototype development
Knowledge representation and data visualization
SEEK core technologies: increasing intelligence about legacy sources
Bio-computing/ evolutionary algorithms

Figure 9.10. Relationship between SEEK approach to knowledge extraction and future research.

5.1 Extending the intelligence of SEEK components regarding legacy source composition

In addition to the general knowledge management challenges articulated by Maybury (2001), we see the following specific technical challenges that must be overcome to build the next-generation SEEK toolkit:

- Technologies to discover the analysis capabilities of the source. This is akin to the field of database integration, where mediators need to know the query and data management capabilities of the underlying data sources (Li et al. 1998). Currently, the SEEK toolkit relies on input from domain users to configure the SEEK analysis engine with information about which type of analysis the source is capable of carrying out natively, and which type(s) of analysis must be done inside the toolkit. A possible solution could be the advertisement of source capabilities using Web service-type descriptions which can be processed by the SEEK toolkit. Initially, these descriptions are provided by domain experts manually. In the long-run, tools are needed to produce these descriptions automatically, a goal shared by the Web services community.

- Technologies to discover the degree of detail of data stored in the source. Although the SEEK toolkit is capable of discovering the database schema of legacy sources, not enough information is available to learn about the level of detail that is available for the accessible data. For example, knowing that a relational table in a database contains sales data for the products sold, does not imply knowing whether this sales data is stored by product, for each day and region, or if it only exists in aggregated form, for example, for all sales for all regions for each year. Extracting this knowledge from the schema itself is difficult; a better approach is to look for database loading and maintenance procedures such as triggers, stored procedures or external load routines, and use semantic analysis (SA) to discover their meaning and side-effects.

In addition, there is a need for metrics to help tune the knowledge extraction system to individual legacy sources. In general, the more accessible and open a legacy source is, the easier and more comprehensive the knowledge extraction process can be. For example, at one end of the spectrum, a legacy source which provides queryable[2] access to the data repository as well as access to all of the application code source files is much easier to work with then a source which provides only limited access to the data (e.g., in the form of periodic data snapshots) and no available application source code. A corporate database accessed by customized application code which is under the control of the party using the SEEK system is an example of the first type. An ERP or MRP system under the control of a third-party is an example of the second. Obviously more work needs to be done in order to develop a reference model that relates source types to the necessary extraction capabilities and/or existing extraction technologies. A formal model will greatly aid in directing future research.

5.2 Advanced approaches to enhance knowledge extraction

Research to test the applicability of and to further develop evolutionary and randomized algorithms has the potential to drastically speed up customization of the knowledge extraction templates and, at the same time, increase the quality of the data extracted. For example, in SEEK, an adaptive search mechanism incorporating biological evolution (e.g., a genetic algorithm), hill-climbing, or randomization (just to name a few) can help reduce the complexity of finding a match between the knowledge extraction templates and the source representation.

[2] *Queryable sources* allow the extraction toolkit to query information at the source, so periodic polling can be used to detect schema and data as well as changes to them.

Issues to be researched include determining the best possible representation of templates to be used by biological and randomized algorithms, how to evaluate the closeness of a template to the real-world representation, how to penalize outliers, etc. Based on preliminary investigations in related application domains (e.g., in configuring data warehouses for decision-support (Lee and Hammer 2001)), we believe it is feasible to use biological algorithms to trade-off a significant amount of search speed for only a slight reduction in the quality of the customized templates. Use of biological models to aid in the customization and fine-tuning of templates could result in a dramatic reduction in the amount of manual effort that has to be invested during the configuration of extraction toolkits.

5.3 Knowledge representation and visual modeling at the user level

Another research area with great potential for improving knowledge extraction is the development of suitable visualization models for displaying domain knowledge at the user interface-level. A suitable model would be an important tool to support creation and customization of the extraction templates, as well as browsing of the query results by the decision makers. For example, based on a few generic (application-specific) seed templates, end-users could potentially view and define new templates to extract previously untapped knowledge or modify existing templates. This reduces the dependency on domain experts and increases usability of the extraction toolkits.

A potential visualization model uses graphical metaphors that are appropriate and personalized for the end-user. By metaphors, we are referring to familiar, everyday processes/objects that could be employed to help users better visualize an unfamiliar or complex phenomenon. For example, this could enable an analyst whose interest is limited to a subset of all the available data to select and visualize the relevant knowledge in a form s/he is accustomed to and can easily understand. Interesting research issues include the development of suitable metaphors as well as representation of the different levels of detail to support drill-down and advanced browsing.

6. Conclusions

This chapter has presented an overview of the basic approaches and research issues associated with the extraction and composition of knowledge in legacy sources. The problems of leveraging knowledge resident in firms to support supply chain coordination are non-trivial. Considerable further research and development is needed to relieve firms' current burden of integrating their legacy systems with decision support applications. Nonetheless, we believe that the SEEK architecture detailed in Section 4.1 is a robust approach to leveraging legacy knowledge, meeting the functional requirements (Section 2.3) for rapid instantiation, connection to heterogeneous sources, and data composition. SEEK also provides a level of security between firms' data and external applications and users, answering queries based on proprietary data while shielding the data itself. Returning to our example of a construction general contractor in Section 2, the SEEK toolkit promises an ability to integrate supply chain information across a network of subcontractors, providing new capabilities to coordinate production schedules in the complex and uncertain environment that characterizes construction projects. More broadly, the SEEK toolkit and approach can be seen as an important enabler, allowing deployment of supply chain decision support tools across business networks of varying size, composition, and sophistication more quickly than is currently possible.

Acknowledgements: The authors would like to thank the faculty (Sherman Bai, Joseph Geunes, Raymond Issa, and Mark Schmalz) and students (Sangeetha Shekar, Nikhil Haldavnekar, Huanqing Lu, Bibo Yang, Jaehyun Choi, and Rodrigo Castro-Raventós) involved on the SEEK project for their contributions to the research and understanding behind this chapter. The authors also thank the National Science Foundation who supported SEEK research under grant numbers CMS-0075407 and CMS-0122193.

References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. (1996). "Fast Discovery of Association Rules." Advances in Knowledge Discovery and Data Mining, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., AAAI/MIT Press.

Ahmed, R., Albert, J., Du, W., Kent, W., Litwin, W., and Shan, M.-C. (1993). "An Overview of Pegasus." Research Issues in Data Engineering, 273-277.

Aiken, P. (1996). Data Reverse Engineering: Slaying the Legacy Dragon, McGraw-Hill.

Aiken, P., Muntz, A., and Richards, R. (1994). "DoD legacy systems: Reverse engineering data requirements." Communications of the ACM, 37(5), 26-41.

Amor, R., and Faraj, I. (2001). "Misconceptions about integrated project databases." ITcon, 6, 57-66.

Arens, Y., Chee, C.Y., Hsu, C., and Knoblock, C. (1993). "Retrieving and Integrating Data from Multiple Information Sources." International Journal of Intelligent & Cooperative Information Systems, 2(2), 127-158.

Ballard, G. (1999). "Improving Work Flow Reliability." Seventh Annual Conference of the International Group for Lean Construction, IGLC-7, Berkeley, CA, July 26-28, 1999, 275-286

Bayardo, R., Bohrer, W., Brice, R., Cichocki, A., Fowler, G., Helal, A., Kashyap, V., Ksiezyk, T., Martin, G., Nodine, M., Rashid, M., Rusinkiewicz, M., Shea, R., Unnikrishnan, C., Unruh, A., and Woelk, D. (1996). "Semantic Integration of Information in Open and Dynamic Environments." MCC-INSL-088-96, MCC.

Beyer, D., and Ward, J. (2002). "Network server supply chain at HP: a case study." Supply Chain Structures: Coordination, Information, and Optimization, J.-S. Song and D. Yao, eds., Kluwer Academic Publishers, Boston/Dordrecht/London, 257-282.

Boulanger, D., and March, S.T. (1989). "An approach to analyzing the information content of existing databases." Database, 1-8.

Cai, Y., Cercone, N., and Han, J. (1991). "Attribute-oriented Induction in Relational Databases." Knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. J. Fawley, eds., AAAI/MIT Press, Cambridge, MA.

Casanova, M.A., and Sa, J.E.A.D. (1983). "Designing entity-relationship schemas for conventional information systems." Third International Conference on Entity-Relationship Approach, August 1983.

Castro-Raventós, R. (2002). "Comparative Case Studies of Subcontractor Information Control Systems," M.S. Thesis, University of Florida.

Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., and Widom, J. (1994). "The TSIMMIS Project: Integration of Heterogeneous Information Sources." Tenth Anniversary Meeting of the Information Processing Society of Japan, Tokyo, Japan, October 1994, 7-18.

Chiang, R.H. (1995). "A knowledge-based system for performing reverse engineering of relational database." Decision Support Systems, 13, 295-312.

Chiang, R.H.L., Barron, T. M., and Storey, V. C. (1994). "Reverse engineering of relational databases: Extraction of an EER model from a relational database." Data and Knowledge Engineering, 12(1), 107-142.

Chikofsky, E.a.J.C. (1990). “Reverse engineering and design recovery: A taxonomy.” IEEE Software, 7(1), 13-17.

Codd, E.F. (1970). “A Relational Model for Large Shared Data Banks.” Communications of the ACM, 13(6), 377-387.

Connolly, D. (1997). “Extensible Markup Language (XML).” W3C.

DARPA. (2000). “The DARPA Agent Markup Language Homepage.”

Davis, K.H. (1995). “August-II: A tool for step-by-step data model reverse engineering.” IEEE Second Working Conference on Reverse Engineering, May 1995, 146-155.

Davis, K.H., and Aiken, P. (2000). “Data reverse engineering: A historical survey.” IEEE Seventh Working Conference on Reverse Engineering, November 2000, 70-78.

Davis, K.H., and Arora, A.K. (1985). “Methodology for translating a conventional file system into an entity-relationship model.” Fourth International Conference on Entity-Relationship Approach, November 1985, 148-159.

Davis, K.H., and Arora, A.K. (1987). “Converting a relational database model into an entity-relationship model.” Sixth International Conference on Entity-Relationship Approach, October 1987, 271-285.

Dayani-Fard, H., and Jurisica, I. (1998). “Reverse engineering: A History - Where we’ve been and what we’ve done.” IEEE Fifth Working Conference on Reverse Engineering, April 1998, 174-182.

DeGroot, M.H., and Schervish, M.J. (2002). Probability and Statistics, Addison-Wesley Publishing, Reading, MA.

Dumpala, S.R., and Arora, S.K. (1981). “Schema translation using the entity-relationship approach.” Second International Conference on the Entity-Relationship Approach, August 1981, 337-356.

Elmasri, R., and S.B. Navathe. (2000). Fundamentals of database systems, 3rd ed., Addison-Wesley.

Englebert, V., and Hainaut, J.-L. (1999). “DB-MAIN: A next generation Meta-CASE.” Information Systems Journal, 24(2), 99-112.

Gallego, G., and Özer, Ö. (2002). “Optimal Use of demand information in supply chain management.” Supply Chain Structures: Coordination, Information, and Optimization, J.-S. Song and D. Yao, eds., Kluwer Academic Publishers, Boston/Dordrecht/London, 119-160.

Genesereth, M., and Duschka, O. (1997). “Answering Recursive Queries Using Views.” ACM Symposium on Principles of Database Systems, Tucson, AZ, June 1997, 109-116.

Hainaut, J.-L. (1991). “Database reverse engineering: Models, techniques, and strategies.” 10th International Conference on Entity-Relationship Approach, November 1991, 729-741.

Hammer, J., Breunig, M., Garcia-Molina, H., Nestorov, S., Vassalos, V., and Yerneni, R. (1997a). "Template-Based Wrappers in the TSIMMIS System." Twenty-Third ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, May 23-25, 1997, 532.

Hammer, J., Garcia-Molina, H., Ireland, K., Papakonstantinou, Y., Ullman, J., and Widom, J. (1995). "Integrating and Accessing Heterogeneous Information Sources in TSIMMIS." AAAI Symposium on Information Gathering, Stanford, CA, March 1995, 61-64.

Hammer, J., McHugh, J., and Garcia-Molina, H. (1997b). "Semistructured Data: The TSIMMIS Experience." First East-European Symposium on Advances in Databases and Information Systems (ADBIS '97), St. Petersburg, Russia, September 1997, 1-8.

Hammer, J., Schmalz, M., O'Brien, W., Shekar, S., and Haldavnekar, N. (2002). "Knowledge Extraction in the SEEK Project." TR-0214, University of Florida, Gainesville, FL 32611-6120, 30.

Han, J., and Kamber, M. (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, CA.

Han, J., Nishio, S., Kawano, H., and Wang, W. (1998). "Generalization-based Data Mining in Object-Oriented Databases Using an Object-cube Model." Data and Knowledge Engineering, 25, 55-97.

Hensley, J., and Davis, K.H. (2000). "Gaining domain knowledge while data reverse engineering: An experience report." Data Reverse Engineering Workshop, EuroRef Seventh Reengineering Forum, January 2000.

Horwitz, S., and Reps, T. (1992). "The use of program dependence graphs in software engineering." Fourteenth International Conference on Software Engineering, Melbourne, Australia, May 1992.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). "Data Clustering: A Survey." ACM Computing Surveys, 31(2), 264-323.

Johannesson, P., and Kalman, K. (1989). "A method for translating relational schemas into conceptual schemas." Eighth International Conference on the Entity-Relationship Approach, November 1989, 271-285.

Kirk, T., Levy, A., Sagiv, J., and Srivastava, D. (1995). "The Information Manifold." AT&T Bell Laboratories,

Klug, A.C. (1980). "Entity-relationship views over uninterpreted enterprise schemas." First International Conference on the Entity-Relationship Approach, August 1980, 39-60.

Lassila, and Swick. (1999). "Resource Description Framework (RDF) Model and Syntax Specification."

Lee, H.L., and Whang, S. (2001). "E-Business and Supply Chain Integration." SGSCMF-W2-2001, Stanford Global Supply Chain Management Forum, 20 pages.

Lee, H.L., and Whang, S. (2002). "Supply chain integration over the internet." Supply Chain Management: Models, Applications, and Research Directions, J. Geunes, P. M. Pardalos, and H. E. Romeijn, eds., Kluwer Academic Publishers, Dordrecht/Boston/London, 3-18.

Lee, M., and Hammer, J. (2001). "Speeding Up Warehouse Physical Design Using A Randomized Algorithm." International Journal of Cooperative Information Systems (IJCIS), 10(3), 327-354.

Levy, A., Rajaraman, A., and Ordille, J.J. (1996). "Querying heterogeneous information sources using source descriptions." International Conference on Very Large Databases, Bombay, India, September 1996, 251-262.

Li, C., Yerneni, R., Vassalos, V., Garcia-Molina, H., Papakonstantinou, Y., Ullman, J. D., and Valiveti, M. (1998). "Capability Based Mediation in TSIMMIS." Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, June 2-4, 1998, 564-566.

Markowitz, V.M. and Makowsky, J.A. (1990). "Identifying extended entity-relationship object structures in relational schemas." IEEE Transactions on Software Engineering, 16(8), 777-790.

Maybury, M. (2001). "Human language technologies for knowledge management: challenges and opportunities." Mitre Corporation Technical Note, Mitre Corporation, Bedford, MA,

Melkanoff, M.A., and Zaniolo, C. (1980). "Decomposition of relations and synthesis of entity-relationship diagrams." First International Conference on the Entity-Relationship Approach, August 1980, 277-294.

Microsoft Corp. (2000). "Microsoft Project 2000 Database Design Diagram."

Mitchell, T. (1997). Machine Learning, McGraw-Hill Science., New York, NY.

Moh, C.-H., E-P. Lim, and W-K. Ng. (2000). "Re-engineering structures from Web documents." ACM International Conference on Digital Libraries 2000, 67-76.

Morey, D., Maybury, M., and Thuraisingham, B., eds. (2000). Knowledge Management: Classic and Contemporary Works. MIT Press, Cambridge, MA.

Navathe, S.B., and Awong, A.M. (1988). "Abstracting relational and hierarchical data with a semantic data model." Entity-Relationship Approach, 305-333.

Nilsson, E.G. (1985). "The translation of COBOL data structures to an entity-relationship type conceptual schema." Fourth International Conference on the Entity-Relationship Approach, November 1985, 170-177.

O'Brien, W.J., Fischer, M.A., and Jucker, J.V. (1995). "An economic view of project coordination." Construction Management and Economics, 13(5), 393-400.

O'Brien, W., and Hammer, J. (2001). "Robust mediation of construction supply chain information." ASCE Specialty Conference on Fully Integrated and Automated Project Processes (FIAPP) in Civil Engineering, Blacksburg, VA, January 23-25, 2002, 415-425.

O'Brien, W.J., Issa, R.R., Hammer, J., Schmalz, M., Geunes, J., and Bai, S. (2002). "SEEK: Accomplishing enterprise information integration across heterogeneous sources." ITcon - Electronic Journal of Information Technology in Construction - Special Edition on Knowledge Management, 7, 101-124.

Palomaeki, A., Wolski, A., Veijalainen, J., and Jokiniemi, J. (1993). "Retrospection on the HERMES PROJECT: Implementation of a Heterogeneous Transaction Management System." IEEE RIDE-International Workshop on Interoperability in Multidatabase Systems, Vienna, Austria, April 1993.

Paul, S., and Prakash, A. (1994). "A Framework for Source Code Search Using Program Patterns." Software Engineering, 20(6), 463-475.

Petit, J.-M., Toumani, F., Boulicaut, J.-F., and Kouloumdjian, J. (1996). "Towards the Reverse Engineering of Denormalized Relational Databases." Twelfth International Conference on Data Engineering (ICDE), New Orleans, LA, February 1996, 218-227.

Rahm, E. and P.A. Bernstein. (2001). "A survey of approaches to automatic schema matching," VLDB Journal: Very Large Data Bases, 10, 334-350.

Schlenoff, C., Gruninger, M., Tissot, F., Valois, J., Lubell, J., and Lee, J. (2000). "The Process Specification Language (PSL): Overview and Version 1.0 Specification." NISTIR 6459, NIST, www.mel.nist.gov/psl/pubs/PSL1.0/paper.doc, 83 pages.

Shanmugasundaram, J., Kiernan, J., Shekita, E., Fan, C., and Funderburk, J. (2001). "XPERANTO: Bridging Relational Technology and XML." IBM Research Report, IBM,

Shao, J., and Pound, C. (1999). "Reverse engineering business rules from legacy system." BT Journal, 17(4).

Shapiro, J.F. (1999). "Bottom-up vs. Top-down approaches to supply chain modeling." Quantitative Models for Supply Chain Management, S. Tayur, R. Ganeshan, and M. Magazine, eds., Kluwer Academic Publishers, Boston/Dordrecht/London, 737-760.

Shapiro, J.F. (2001). Modeling the Supply Chain, Duxbury, Pacific Grove, CA.

Silva, A.M., and Melkanoff, A. (1981). "A method for helping discover the dependencies of a relation." Conference on Advances in Database Theory, Toulouse, France, 115-133.

Smith, J.M., Bernstein, P.A., Goodman, N., Dayal, U., Landers, T., Lin, K.W.T., and Wong, E. (1981). "MULTIBASE–Integrating Heterogeneous Distributed Database Systems." National Computer Conference, March 1981, 487–499.

Song, I.-Y., and Froehlich, K. (1995). "Entity-relationship modeling." IEEE Potentials, 13(5), 29-34.

Suciu, D. (1998). "An Overview of Semistructured Data." SIGACT News, 29(4), 28-38.

Tan, G.W., Shaw, M.J., and Fulkerson, W. (2000). "Web-based global supply chain management." Handbook on Electronic Commerce, M. J. Shaw, R. Blanning, T. Strader, and A. Whinston, eds., Springer-Verlag, Berlin, 457-480.

Tayur, S., Ganeshan, R., and Magazine, M., eds. (1999). Quantitative Models for Supply Chain Management. Kluwer Academic Publishers, Boston/Dordecht/London.

Templeton, T., and Al., E. (1987). "Mermaid: A Front-End to Distributed Heterogeneous Databases." Computer Science International Conference on Database Engineering, 695-708.

Tsay, A.A., Nahmias, S., and Agrawal, N. (1999). "Modeling supply chain contracts: a review." Quantitative Models for Supply Chain Management, S. Tayur, R. Ganeshan, and M. Magazine, eds., Kluwer Academic Publishers, Boston/Dordrecht/London, 299-336.

Verdicchio, M., and Colombetti, M. (2002). "Commitments for agent-based supply chain management." SIGecom Exchanges, 13-23.

Weiss, S.M., and Kulikowski, C.A. (1991). Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems, Morgan Kaufmann, San Mateo, CA.

Wiggerts, T., Bosma, H., and Fielt, E. (1997). "Scenarios for the identification of objects in legacy systems." IEEE Fourth Working Conference on Reverse Engineering, September 1997, 24-32.

Wills, L.M. (1994). "Using attributed flow graph parsing to recognize clichés in programs." International Workshop on Graph Grammars and Their Application to Computer Science., November 1994, 101-106.

Workflow Management Coalition. (2002). "Workflow process definition interface – XML process definition language, version 1.0." WFMC C-1025, The Workflow Management Coalition, `http://www.wfmc.org/standards/docs/TC-1025_10_xpdl_102502.pdf`, 87 pages.

Xyleme, L. (2001). “A dynamic warehouse for XML Data of the Web.” IEEE Data Engineering Bulletin, 24(2), 40-47.

Zamanian, M.K., and Pittman, J.H. (1999). “A software industry perspective on AEC information models for distributed collaboration.” Automation in Construction, 8, 237-248.

Chapter 10

COLLABORATION TECHNOLOGIES FOR SUPPORTING E-SUPPLY CHAIN MANAGEMENT

Stanley Y.W. Su, Herman Lam, Rakesh Lodha
Department of Computer & Information Science & Engineering
University of Florida
{su, hlam, rlodha}@cise.ufl.edu

Sherman Bai, Zuo-Jun (Max) Shen
Department of Industrial Systems Engineering
University of Florida
{bai, shen}@ise.ufl.edu

Abstract This chapter presents the design and implementation of an Event-Trigger-Rule-based electronic supply-chain management system (ESCM). The ESCM is constructed by a network of Knowledge Web Servers, each of which consists of a Web server, an Event Manager, an Event-Trigger-Rule (ETR) Server, a Knowledge Profile Manager, a Persistent Object Manager, a Metadata Manager, a Negotiation Server, and a Cost-Benefit Evaluation Server. Together, they manage the activities and interactions among Manufacturers, Distributors and Retailers. ESCM offers a number of features. First and foremost is the flexibility offered to business entities in defining their own rules according to their own business strategies. Second, since the rules that control the business activities are installed and processed by the multiple copies of the ETR server installed at business entities' individual sites, their privacy and security are safeguarded. Third, ESCM's event, event filtering and event notification mechanisms keep both Buyers and Suppliers better informed with more timely information about business events so that they or their software systems can take the proper actions in different business processes.

1. Introduction

A supply chain is comprised of an inter-linked network of suppliers, manufacturers, distributors, retailers and customers. In a typical supply chain, materials flow from the suppliers through manufacturers and distributors to the retailers and customers, while demand information flows in the opposite direction. Supply chain management deals with the problem of how to efficiently integrate the decisions made by different members in the supply chain so as to minimize system-wide costs subject to certain service requirements.

With the current development of Internet and information technologies, many companies have engaged in what we might call electronic supply chain management (ESCM), in which many traditional supply chain management activities are carried out electronically. The current web and distributed object technologies provide a basic *information infrastructure* over the Internet to *interconnect* enterprises and allow data from distributed application systems to be shared among members in a supply chain. However, there are still some limitations in the basic information infrastructure when it is used for collaborative ESCM. We shall discuss these in the following paragraphs.

First, for companies across a supply chain to coordinate their product, financial and information flows, they must have access to **accurate** and **timely** information reflecting the status of these flows. Information sharing is the most effective way to counter the problem of demand information distortion in a supply chain – the well known "bullwhip effect" (Lee, Padmanabhan, and Whang (1997)). Information distortion often arises when (i) partners make use of local information to make demand forecasts and pass the forecasts to upstream partners; (ii) partners make ordering decisions based on local economic factors, local constraints or performance measures; and (iii) gaming behaviors exaggerate orders when there are perceived uncertainties in supply conditions. These distortions are amplified from one level to another in a supply chain, and are considered to be some of the primary causes of inefficiencies in a supply chain.

One limitation of the current basic information infrastructure is that whenever some data is required upstream or downstream in a supply chain, it has to be manually searched or queried. In this manner, important data may not be delivered to the right supply chain stage in a timely manner. The Internet is based on the client-server paradigm of computing. Servers on the Internet are initially isolated from each other and only respond to requests that are given to them. This pull model of data access requires a system that is doing the pulling to know what to pull and when to pull for information. A supply chain management infrastructure should guarantee that such information is accessible by the appropriate parties on a timely basis. Also, while sharing information, the secrecy of each partner's confidential data has to be maintained.

A second problem of the existing infrastructure is that it only enables the establishment of a **data** network instead of a **knowledge** network. There is no facility on the Internet for expressing human/organizational knowledge in the form of computer-processible rules, which can be used to automatically extract meaningful data, activate application systems to process the data, notify relevant people about the occurrences of events in a timely manner, and enforce data integrity and security constraints, business policies and strategies, government regulations, etc. Knowledge captures the successful experiences of people and enterprises and helps to automate business processes.

The third problem of the existing infrastructure is that it does not provide e-services that are directly relevant to e-business in general and e-supply chain in particular. E-services which support business rule management, business constraint processing, cost-benefit evaluation and business negotiation are needed as middleware services within the infrastructure.

In this work, we have developed an information infrastructure, which is enhanced by a number of collaboration technologies and their corresponding servers developed at the Database Systems Research & Development Center at the University of Florida: namely, Event-Trigger-Rule (ETR) Server, Constraint Satisfaction Processing Server, Cost-benefit Evaluation Server, and Automated Negotiation Server. The e-services provided by these servers are used to support the interaction and collaboration of business entities in an e-supply chain. The ETR technology combines the features of the push and pull models of information access and delivery, and provides more advanced event filtering and composite event processing capabilities. It is used to capture and share the knowledge of business enterprises in supply chains. Business knowledge is represented by high-level specifications of business events, triggers, rules, and processes that are associated with distributed objects.

The inclusion of events, triggers, and rules in object class definitions is an extension to the traditional distributed object model and makes distributed objects "active" (i.e., active distributed objects) in that an activation of an operation (or method) of a distributed object can post events to automatically trigger the processing of business rules. In a supply chain environment, there are many situations in which one business process depends on another business process' result. Rules can be used to specify and conditionally activate business processes based on the result of another process. Business entities can also define their business strategies, regulations, and constraints in terms of rules. Since rules are high-level specifications rather than code in application programs, they are easier for users or the persons who define the rules to understand and modify.

In addition to the ETR Server for managing business events, triggers and rules, we believe that additional servers and tools are needed to build a powerful information infrastructure for supporting supply chain management. For example, when a Buyer receives quotes from Suppliers, these quotes need to be evaluated in order to first determine if they satisfy the constraints and requirements of the Buyer, and then, if multiple quotes satisfy the constraints and requirements, find the most cost-beneficial one. Thus, a server for constraint satisfaction processing is needed to compare product requirements and constraints of Buyers and Suppliers, and a server is also needed to perform cost-benefit evaluation of quotes or business contracts. Also a Buyer may want to negotiate with a Supplier over the price, quantity, delivery date, and/or attributes of a product or service. For example, a delivery truck's accident may require re-negotiation of a new delivery schedule, which may involve negotiations over the schedules up and down a supply chain. These negotiations are not just one-on-one negotiations. They may involve negotiations with multiple parties simultaneously (i.e., multilateral negotiations). It would be beneficial if a part or all of the negotiation tasks could be carried out automatically over the Internet as a part of e-business activities. Thus, the services of a Negotiation Server are needed to perform automated negotiations. A tool for capturing the product/service constraints of both Buyers and Suppliers is also needed for the analysis of negotiation proposals and the generation of counter-proposals.

In a supply chain management system, these servers may depend on each other for services. For example, when a Buyer receives multiple quotes from Suppliers (an event); the ETR server is notified of the event, which triggers the Cost-Benefit Evaluation Server to evaluate the quotes to perform supplier selection. The Buyer may want to negotiate with a selected Supplier for the terms and conditions of the purchase. In such cases, a Negotiation Server installed at the Buyer site will be activated to perform automated negotiation with the Negotiation Server installed at the Supplier's site. When a proposal is received by the negotiation server of the Buyer/Supplier, the services of a Constraint Satisfaction Processing Server can be used to evaluate the proposal against the constraints and requirements of the Buyer/Supplier. During the negotiation process, if the conditions specified in a negotiation proposal do not satisfy the constraints of the Buyer/Supplier, a rule that implements a negotiation strategy, which is managed by the ETR Server of the Buyer/Supplier, can be triggered to allow some concession before a counter-proposal is generated and sent to the Supplier/Buyer. Proposals and counter-proposals can be sent back and forth between the two Negotiation Servers until either an agreement is reached by both negotiation parties or one party unilaterally terminates the negotiation process. The acceptance and termination of a negotiation can again be determined by rules.

To meet the requirements discussed above, we have integrated an ETR Server (Lee (2000); Lee, Su, and Lam (2001); Su and Lam (2000)), a Cost-Benefit Evaluation Server (Su, Dujmovic, Batory, Navathe, and Elnicki (1987); Yu (2001); Liu, Yu, Su, and Lam (2002)), a Negotiation Server (Huang (2000); Hammer et al. (2000); Su, Huang, and Hammer (2000); Su et al. (2001); Li (2001)), a Timer Event Server, a Constraint Satisfaction Processing (CSP) Server (Huang (2000)) and their associated GUI tools developed at the Database Systems Research & Development Center to form an enhanced information infrastructure for supply chain management. We have also developed a number of supply chain scenarios to demonstrate how these servers can be used to support supply chain management. The developed scenarios involve three business entities: a Retailer, a Distributor and a Manufacturer. By using these three entities, we are able to test and demonstrate the interactions of one business entity with two adjacent neighbors. The same business activities and interactions can occur in any segment of a supply chain. Thus, the results of this work can be applied to a supply chain with any number of business entities.

The major contributions of this work are: 1) the integration of several existing servers and tools to construct an ESCM, 2) the development and implementation of several supply chain scenarios by using the services of the integrated servers and tools, and 3) the testing and demonstration of the developed ESCM system. The intended contributions summarized in terms of system features and functions will be provided in the conclusion section.

This chapter is organized as follows. First, we explain in brief the ESCM model and the scenarios we have developed to demonstrate the utilities of the servers in the proposed information infrastructure. Then, we describe the architecture of ESCM and different technologies used in its implementation. Finally, we conclude by pointing out the benefits of the integrated system and propose issues for further research.

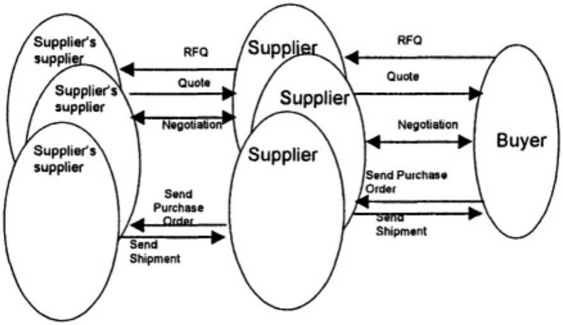


Figure 10.1. ESCM Model.

2. ESCM MODEL AND SCENARIOS

In order to demonstrate the use of ESCM infrastructure technologies for e-supply chain management, we developed a generic model for modeling the interactions among three business entities that form two links of a supply chain. The generic model is then used to develop a number of generic scenarios for the prototype implementation of an ESCM system.

2.1 Three Classes of e-Business Applications

The ESCM model consists of three business entities acting as Buyers, Suppliers and Suppliers' Suppliers and their interactions. There can be many Buyers interacting with the same Supplier. Figure 10.1 shows this abstract model, which is independent of the technology used to implement it.

In this model, a Buyer sends a Request-for-Quote (RFQ) to one or many Suppliers. The RFQ contains the requirements of the Buyer such as product name, product number, quantity of product, delivery date, etc. Upon receiving the RFQ, each Supplier checks its business policy. The Supplier also checks its inventory. If inventory is not enough to meet the demand, then it in turn sends an RFQ to its Suppliers (i.e., the Supplier's Suppliers). A Supplier's Supplier sends its quote to the Supplier and, based on this quote's information, the Supplier sends its quote to the Buyer. The Buyer then checks different quotes received from different Suppliers and does a cost-benefit evaluation on each of them. After this cost-benefit evaluation, if no quote matches with the Buyer's expectations, then it can start negotiations with the Suppliers simultaneously (i.e., the first level negotiations). Each Supplier can check whether it has sent a request-for-quote to its suppliers and, and based on the returned quotes, it may start negotiations with its suppliers (i.e., the second level negotiations). Once the second level negotiations are completed successfully, the Supplier can use the negotiated results to continue its negotiation with the Buyer. The Buyer/Supplier can define its own business rules used in the negotiations. If a negotiation is unsuccessful, then the Buyer can resend a modified RFQ to the Suppliers and the whole process can start all over again. Upon a successful negotiation, the Buyer sends a purchase order to a respective Supplier. The Supplier in turn sends a purchase order to its respective Supplier. Upon receiving the shipment from the Supplier's Supplier, the Supplier then sends the product to the Buyer.

The above model represents two links (three business entities) of a supply chain. It shows that the business interactions between two entities in one link can affect the interactions between two entities in another link (e.g., the result of a negotiation in one link can affect the negotiation of the other link.) The business interactions shown in the model can similarly occur in other segments of a supply chain.

2.2 ESCM Scenarios

To verify the effectiveness of the ESCM model for e-supply chain applications, we have developed a set of scenarios for prototype implementations Lodha (2002). The purposes of these scenarios are as follows:

1 To demonstrate a prototype implementation that involves the integration of inter-enterprise business functions within the scope of the ESCM model.

2 To demonstrate how different technologies such as an Event-Trigger-Rule technology, an automated negotiation technology, a cost-benefit evaluation technology, and a constraint satisfaction processing technology can be integrated to support a supply chain management system.

These scenarios include major players of supply chains, such as Retailers, Distributors, and Manufacturers. They describe business processes, which include determining qualified suppliers, requesting quotes, and automating negotiations between any two entities in the supply chain. They also include inventory replenishment, quote evaluation, transport decisions, cost-benefit evaluation, and order processing for any entity in the chain. These scenarios do not show what specific software or technology is used to implement them. Different supply chain entities can use different tools, systems and methodologies to implement the scenarios and different approaches to solve the supply chain problems.

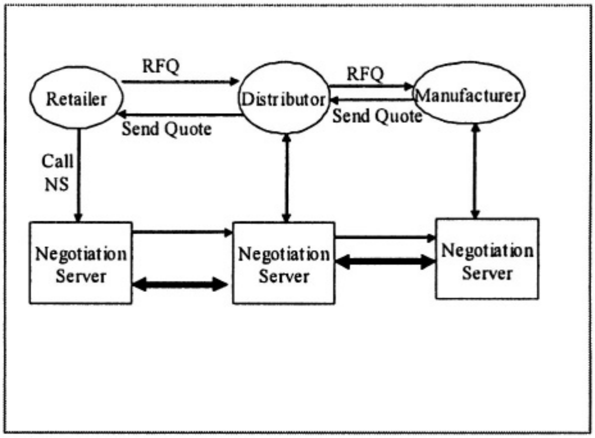


Figure 10.2. A Two-Tier Scenario.

In this chapter, we describe one of the scenarios: a two-tier scenario shown in Figure 10.2. In this scenario, the Retailer first sends the Distributor an RFQ for the purchase of some units of a product. The Distributor checks its inventory and finds that it cannot satisfy the quantity requested by the Retailer. It thus sends an RFQ to the Manufacturer for the purchase of sufficient units of the product in order to satisfy the Retailer's order. Upon the receipt of the quote from the Manufacturer, it generates and sends a quote to the Retailer. The Retailer checks the quote and finds that some conditions given in the quote are not satisfactory (e.g., the price is too high, the delivery date is too late, etc.). The retailer then begins a negotiation process with the Distributor to address these conditions. The Negotiation Server, which acts on behalf of the Retailer, sends a proposal to the Negotiation Server of the Distributor. Upon receiving the proposal from the Retailer, the Distributor's Negotiation Server checks the proposal against the Distributor's pre-registered policies and constraints and realizes that it in turn needs to start a negotiation process with the Manufacturer in an attempt to satisfy the Retailer's demand. In this case, the negotiation with the Retailer would have to depend on the outcome of the negotiation with the Manufacturer. The Distributor has to temporarily suspend its negotiation with the Retailer and to complete the negotiation with the Manufacturer first.

In each of the two bi-lateral negotiation processes described above, negotiation proposals and counter-proposals can be sent between a pair of business entities until either an agreement is reached or one party unilaterally terminates the negotiation process. The allowable number of proposal exchanges can be controlled by each side using a termination rule or a time-out mechanism.

There are many other scenarios in a supply chain in which the automated negotiation service is needed. For example, while in the middle of a product delivery, if the Distributor's transportation truck has an accident, then the delivery will not be completed on the specified date/time; hence a new delivery date/time is to be negotiated. If the Manufacturer's machinery stops because of a mechanical problem that causes a production delay, then the Manufacturer will have to negotiate with the Distributor on a new delivery date, which may in turn invoke some simultaneous renegotiations with Retailers.

We point out here that the bi-lateral negotiation described in the above scenario is not sufficient when multiple buyers/suppliers are involved. In those situations, multilateral automated negotiation will be needed. Multilateral negotiation requires the ability to perform simultaneous negotiations with multiple business entities, and to take the proper action or to make the proper decision in one negotiation based upon the status/results of the other simultaneous negotiations. Multilateral negotiation is, however, outside of the scope of our current research effort.

3. Modeling of Inter-enterprise Resources

Having presented the ESCM model and an example supply chain scenario, we now explain how inter-enterprise resources are modeled and accessed. In a collaborative e-business environment, it is necessary to have a uniform way to model different types of inter-enterprise resources. The existing distributed object technology exemplified by OMG's CORBA, Microsoft's COM/DCOM and Enterprise Java Beans encapsulates these resources uniformly as distributed objects (DOs) to facilitate their interoperability. However, DOs are "passive" in the sense that they only respond to method calls from other DOs. They cannot "actively" enact business processes or perform operations to enforce business constraints, policies, and regulations in response to some business events. In this work, we use an active distributed object (ADO) technology (Lam and Su (1998); Lee (2000)) developed at the Database R&D Center of the University of Florida, which also incorporates business events, rules, and processes in the modeling and processing of various types of enterprise resources and provides GUI tools (Parui (1999)) for capturing the active properties of ADOs. The approach is a generalization of the Java Beans approach to link loosely coupled components together. The difference is that, in ADOs, components are not limited to graphical components and the linking of these components is done at a high level using rule and process specifications (as opposed to coding in a programming language).

3.1 Active Object Model

To give distributed objects their active properties, the traditional object model has been extended into an active object model (or AOM). In AOM, objects of a class are modeled in terms of attributes and methods, such as the traditional object model. Additionally, events, triggers, and action-oriented rules can be optionally included in an object class specification. Just as events, triggers, and action-rules can transform a conventional database system into an active database system (Dayal (1998), Chakravarthy, Anwar, Maugis, and Mishra (1994), Widom (1996)), they are used in AOM to transform distributed objects (DOs) into active distributed objects (ADOs). In effect, methods are the "incoming" application program interfaces (APIs) of an object (i.e., they define the functions that can be performed by the object); and events are "outgoing" APIs (i.e., they define the signals that can be generated by the object).

There is a trend for supply chain entities to explicitly specify data conditions, events of interest that may occur in a supply chain environment and business policies, constraints, regulations, strategies, etc., in terms of business events and business rules. Some of these events and rules (i.e., public events and rules) become important inter-enterprise resources that need to be shared across collaborative business entities. Others are private events and rules applicable only to a business entity. In the following subsections, we shall explain how events, rules and triggers are defined and how they are used to make distributed objects active.

3.1.1 Events. Generally speaking, events can be data conditions and/or things that happen in the Internet, which need to be monitored. For example, new data is made available on a site, a web page has been modified, an application system is about to be invoked, a user strikes a key on the keyboard, a signal is received from an external device, etc. Users or software systems can subscribe to events and be notified when events occur. Events can have parameters (i.e., parameterized events), which are data associated with the events that are to be transmitted through event notifications.

In the context of a supply chain, events can be used to represent things that happen during different business processes. For example, a Supplier would like to be notified when a Buyer posts a Request-for-

Quote (RFQ) event. This Supplier is registered with the Buyer's site so that it can receive the RFQ for a particular product. The Supplier can specify different products of interest by providing values to certain attributes, which are specified in an event filter template. Similarly, the Buyer would be interested in being notified when the Supplier posts a Send-Quote event. Another point of interest would be when a certain amount of time has elapsed after sending the RFQ. When this point of interest occurs, the Buyer may specify some actions to be taken at this point depending on its business strategy.

3.1.2 Rules. In AOM, rules are Condition-Action-Alternative Action (CAA) rules, which can be used to define business constraints, policies, regulations, and integrity and security constraints. Each CAA rule represents a small granule of control and logic. A number of related rules can form a rule structure to express a larger granule of control and logic for modeling a more complex policy, regulation, etc. A rule can participate in multiple rule structures, thus making each rule reusable. A rule has a rule name and can have parameters. When the rule is invoked upon the occurrence of some event, it evaluates the CONDITION clause of the rule. If the result is true, the operations specified in the ACTION clause are executed. Otherwise, the operations specified in the ALTACTION clause are executed.

A rule also has a RULEVAR clause, which allows variables to be defined in a rule. The variables can be persistent or temporary. It is also possible to define customizable rule variables, which can be assigned different values for different users, making the rule a "parameterized rule".

Different from the Event-Condition-Action rules used in some active database management systems (Chakravarthy, Anwar, Maugis, and Mishra (1994); Dayal (1998); Stonebraker, Hanson, and Potamianos (1988); Su and Yu (1997); Widom (1996)), events and rules in AOM are separately defined by users or business organizations; and events are tied to rules by trigger specifications (to be described next). This separation is important in a distributed environment in which software systems are loosely coupled. In a supply chain management system, different business entities define and post events, which are subscribed to by other entities. These entities may define different rules for implementing different business strategies. For example, every time a Supplier receives an RFQ event, a rule may be triggered to check the quantity of the requested quote; if the quantity is too small, then the Supplier notifies the Buyer that the Supplier is not interested in the RFQ because the quantity is too small. These rules are managed and used by an Event server and an ETR server installed at the subscribers' sites; thus subscribers' security and privacy will not be compromised.

3.1.3 Triggers.

A trigger relates a structure of events with a structure of rules. An event structure has two parts: namely, a TRIGGEREVENT and an EVENTHISTORY. The TRIGGEREVENT part specifies a number of alternative events. The occurrence (or posting) of any one of these events would initiate the evaluation of the EVENTHISTORY part. The EVENTHISTORY part can be a complex event expression, which makes reference to some events that have already occurred. "E1 follows E2 follows E8," and "E4 and E7 occurred within a certain time window" are examples of event history expressions. Thus, upon the occurrence of an event specified in a TRIGGEREVENT expression, the EVENTHISTORY expression is evaluated. If the result is true, then the rule structure given in the trigger specification is processed. Triggers can be defined by business entities and make reference to events and rules defined by other entities.

3.2 Active Distributed Object (ADO)

The event-trigger-rule specification in the AOM provides an expressive modeling language/tool for capturing the semantics of heterogeneous resources more completely as ADOs. ADOs not only encapsulate these resources as distributed objects but also allow business rules to be triggered to enforce security and integrity constraints, business policies and strategies, or regulations. The ETR Service, which we later describe in Section 5.1, is used for implementing the ADO technology to support interoperability among heterogeneous resources.

Events can be defined and posted at different times relative to the execution of a method. For example, before and after the execution of a method, a Before-method event and an After_method event can be defined and posted to signal that the method is about to be executed and has been executed, respectively. These events provide the "hooks" to trigger a set of rules, which may in turn enact processes or activate application systems. The actions of these rules can be used to customize the behavior of a server object (to satisfy some integrity constraints or e-supply chain management (ESCM) business policies) without having to change the implementation of the component system. This is especially important for the integration of ESCM's component systems, for which we may not have control over the source code. Using the ADO technology, loosely coupled systems can inter-operate together collaboratively with minimal dependency (by publishing and subscribing to events) and at a high level (specification of triggers and rules vs. hard coding).

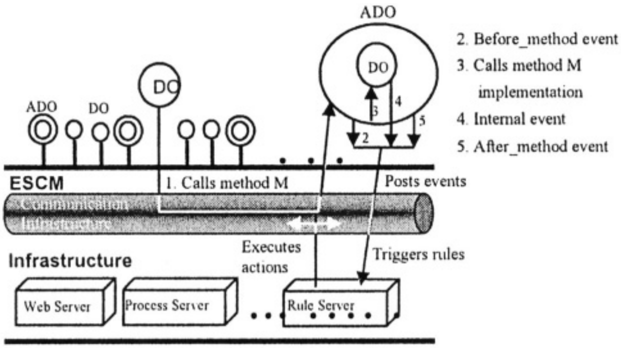
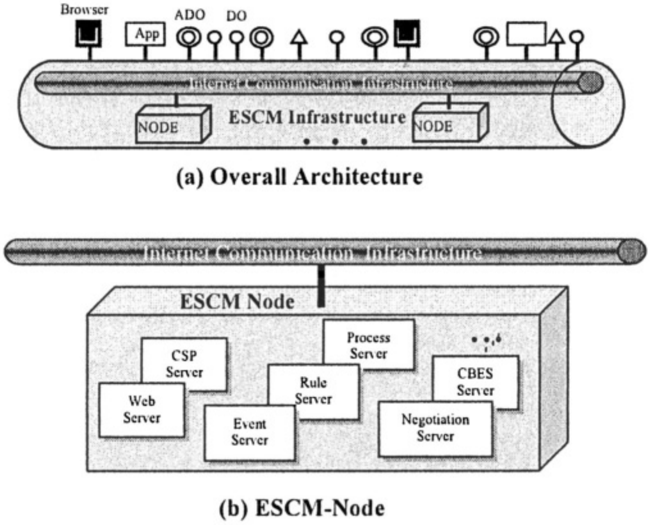


Figure 10.3. Inter-operating Active Distributed Objects.

Figure 10.3 shows a scenario of inter-operating distributed object (DOs), some of which are ADOs. In the scenario, a client object calls a method M of a server object (label 1). Assuming the server object is an ADO, then method M is modified at the compilation time to include a number of event-posting statements. Before the original implementation of method M is executed, a Before_method event (if it exists) is posted (label 2) to trigger a structure of ESCM business rules (a pre-condition to the method). Then the implementation of method M is executed (label 3). During the execution of method M, other internal events (e.g., an exceptional condition) can be posted (label 4) to trigger rules. After the method has completed its execution, an After_method event (if it exists) can be posted (label 5) to trigger another structure of ESCM business rules (post-condition). Thus, the interactions among distributed objects can be done not only through remote method invocations but also through events, and are subject to the control of business rules.

4. ESCM Architecture

The present Web, Internet and distributed object technologies provide a basic infrastructure for interconnecting enterprises and allowing data and application systems to be shared among customers, suppliers, and business partners. In this work, we enhance the basic infrastructure to also provide a number of e-services needed for collaborative-business and other general distributed applications; that is,



(a) Overall Architecture

(b) ESCM-Node

Figure 10.4. ESCM Information Infrastructure.

ESCM Infrastructure = Web/Internet + Existing Distributed Object Technology + Collaborative e-services

The proposed infrastructure and its relationship with existing application systems, distributed objects, agents, Web browsers, and Web servers are depicted in Figure 10.4(a). As shown in Figure 10.4(b), the ESCM Node consists of a Knowledge Web Server (Lee (2000); Lee, Su, and Lam (2001)), a Negotiation Server (Huang (2000); Su et al. (2001)) and a Cost-Benefit Evaluation Server (Yu (2001); Liu, Yu, Su, and Lam (2002)). The Knowledge Web Server consists of a Web server and a number of additional components. These components are: an Event Server, an Event-Trigger-Rule (or ETR) Server, a Knowledge Profile Manager, a Persistent Object Manager (Shenoy (2001)), and a Metadata Manager. These servers form an ESCM Node. The ESCM infrastructure is a network of Nodes deployed at the sites of participating enterprises, much the same way as Web servers are deployed to provide Web services.

5. ESCM Technologies and Servers

In this section, we present a number of core technologies and briefly describe their implementations. Some of our core technologies and servers are not covered in this chapter. Interested readers are referred to the references that we shall provide for additional details.

5.1 Event-Trigger-Rule Technology and Server

An Event-Trigger-Rule (or ETR) server has been implemented in Java (Lee (2000); Lee, Su, and Lam (2001)). Figure 10.5 shows the component architecture of the ETR Server. Each circle shows the Java class name of the module. The classes on the left show the interfaces currently supported for the communication infrastructure. RMI, OrbixWeb, and Vitria Communicator are supported by their interface classes. The ETRMain class identifies the communication infrastructure and hands the information to the ETRServerInterface class, which creates the relevant interface class and exposes the main APIs of the ETR Server to various interfaces. The CLI (Call Level Interface) enables the administrator of the ETR Server to interact with a text menu displayed on the screen. The RuleObjectManager is the entry point to the core ETR Server classes. There are two hash tables: the event hash table and the trigger hash table. The event hash table stores the event to trigger hashing information. By hashing with the event name, the corresponding trigger can be identified. The trigger hash table stores the mapping from the trigger to its triggering events. This makes it possible to edit the triggers and modify the event hash table to make some event entries no longer pointing to the trigger. The Rule Scheduler can perform the scheduling of multiple rules for execution. This includes a special data structure used by the scheduling algorithm. The rules are also executed by threads in order to support efficient processing within a trigger. The RGCoordinator manages the rule group information, of the details of which are beyond the scope of this chapter.

5.2 Constraint Satisfaction Processing Technology and Server

Many problems encountered in e-business such as supplier selection, negotiation proposal evaluation, etc., can be represented as constraint satisfaction processing (CSP) problems. For example, the specifications of products offered by suppliers can be treated as constraint specifications (i.e., products' characteristics, costs, delivery date, etc.) and the product requirements of a buyer can also be treated as a constraint specification, which is to be matched against those of the suppliers in a supplier selection process. A negotiation proposal of a buyer/supplier can also be treated as a constraint specification, which is to be matched with the capabilities/requirements of a supplier/buyer to determine if

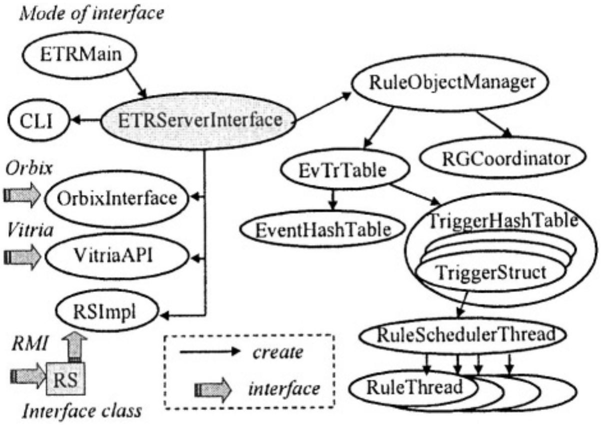
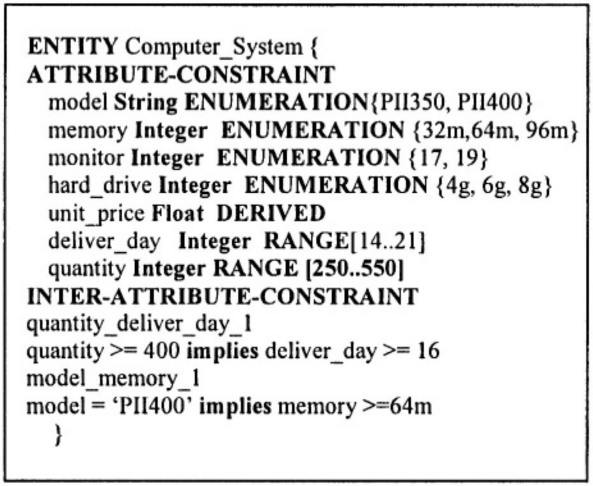


Figure 10.5. Component Architecture of the ETR Server.

the proposal is acceptable or if a counter-proposal should be generated. It is therefore useful to provide constraint satisfaction processing as an e-service.

To provide this e-service, a general object-oriented constraint specification language will be needed. Such a language must be powerful and flexible enough for a buyer/supplier to specify the alternative attribute values that it requires/offers (i.e., attribute constraints) and the inter-relationship between or among these attribute values (i.e., inter-attribute constraints). For example, as shown in Figure 10.6, a supplier of computers may use the following specification to let the potential buyers know about the alternative products it supplies. The attribute types, RANGE and ENUMERATION, are used to specify the alternative attribute values. Logical implication (or If-Then logic) is used to specify inter-attribute constraints. The value for the derived attribute Unit_price is to be derived based on the specific values selected by the buyer.

When an automated system, which represents the buyer, receives the specification shown in Figure 10.6, it would match this against the buyer's product requirements and constraints, which are also specified using the same language constructs. If there are conflicts, the system would have to determine what constraints are violated and the quantitative measure of match or mismatch. This constraint satisfaction problem is not trivial for the following reasons. First, since ENUMERATION and RANGE are used to specify attribute constraints, the comparison of the buyer's and supplier's corresponding attribute constraints would



```
ENTITY Computer_System {
ATTRIBUTE-CONSTRAINT
   model String ENUMERATION{PII350, PII400}
   memory Integer  ENUMERATION {32m,64m, 96m}
   monitor Integer  ENUMERATION {17, 19}
   hard_drive Integer  ENUMERATION {4g, 6g, 8g}
   unit_price Float  DERIVED
   deliver_day  Integer  RANGE[14..21]
   quantity Integer RANGE [250..550]
INTER-ATTRIBUTE-CONSTRAINT
quantity_deliver_day_1
quantity >= 400 implies deliver_day >= 16
model_memory_1
model = 'PII400' implies memory >=64m
   }
```

Figure 10.6. An example constraint specification.

involve set- and range-oriented processing, which can be rather time-consuming. Second, and worse still, after the evaluation of attribute constraints, each attribute may have multiple values that satisfy both the buyer's and supplier's specifications. A potentially large number of combinations of these values will have to be evaluated to see if they satisfy the inter-attribute constraints. The traditional CSP algorithm (e.g., the one used in the commercial system ILOG) would generate all combinations of attribute values that satisfy the attribute constraints of both sides, and evaluate the inter-attribute constraints over these combinations. For example, if there are four Integer type attributes and each attribute constraint defines a range having 1000 acceptable values, e.g., [1000 .. 2000], [5000 .. 6000]. The total number of combinations generated will be 1000*1000*1000*1000=1 trillion. Obviously, testing these combinations against the inter-attribute constraints would consume a huge amount of computing resources. Efficient algorithms for attribute and inter-attribute constraint evaluations are needed.

Third, after checking attribute constraints and inter-attribute constraints, and eliminating those attribute values that violate the constraints of both sides, there may still be several value combinations to be considered. A cost-benefit evaluation of these combinations needs to be performed to quantitatively determine the best alternative (e.g., in a negotiation or supplier selection situation).

Fourth, if a constraint is found to be in violation, an automated system may want to relax its own constraint to satisfy the other side's constraint (e.g., in a bargaining situation), or call the attention of the user it represents. It is therefore important for a constraint satisfaction processor to identify the constraint that is violated. Unfortunately, the traditional CSP systems cannot do this. They can only report if a given specification satisfies the set of constraints or not.

We have developed two efficient algorithms for constraint satisfaction processing to handle the kind of constraint specification described above, and we have implemented a CSP server, which is capable of processing attribute and inter-attribute constraints and identifying the specific constraints that have been violated. Interested readers should refer to Huang (2000) and Su et al. (2001) for the implementation details.

5.3 Cost-benefit Evaluation and Selection Technology and Server

In all kinds of decision-making situations in general, and in business decisions in particular, an individual or a business is often asked to make a choice from a number of presented alternatives. Arriving at a decision involves evaluation of these alternatives from the cost and benefit point of view and then selecting the one that is most beneficial. For example, a company receives a contract proposal with a number of alternative conditions or it receives multiple bids from suppliers. The alternative conditions or the competing bids need to be evaluated to determine their costs and benefits to the company.

A systematic, quantitative, and justifiable cost-benefit evaluation model is needed for this type of evaluation and selection; and a server needs to be implemented to provide this e-service. In previous work, we have developed a cost-benefit decision model (CBDM) for the evaluation and selection of database management systems (Su, Dujmovic, Batory, Navathe, and Elnicki (1987)). In our model, the contents of each product specification are divided into two structures: (1) the cost structure, which contains all the attributes to which costs can be assigned (e.g., a specific disk, an additional memory board, etc.) and (2) the preference structure, which contains all the attributes to which preference scores can be assigned subjectively (note: these two sets of attributes may overlap). The structures are separately analyzed to obtain two aggregated values: an aggregated cost value and the global preference score. These two values are then combined to derive a global cost-preference indicator for the specification of a system under evaluation.

In the preference analysis based on the preference structure, preference scores ranging from zero to one are assigned to all attribute-value pairs using a set of "elementary criteria." An elementary criterion is a mapping from an attribute-value or an attribute-value-range pair to a real number between zero and one. This real value expresses a client's degree of satisfaction with the particular attribute value. For example, if Vendor-Service is an attribute of the data entity Computer, a client may assign a preference score of 0 if the value of Vendor-Service is "mornings only", 0.3 if it is "day time only", and 1 if there is "24-hour service." Here, the score of 1 implies one hundred percent satisfaction. For attributes whose values cannot be enumerated, e.g., Mean-time-to-failure of a disk, evaluation functions can be defined and used as the elementary criteria. The elementary preference scores are weighted and aggregated into a global preference rating using a spectrum of "preference aggregation functions," which are derived from a weighted power mean introduced by Dujmovic (1975):

$$E = (w_1 \cdot e_1^r + w_2 \cdot e_2^r + \ldots + w_n \cdot e_n^r)^{1/r} .$$

By varying the value of r, a spectrum of aggregation functions is generated, including functions such as min, max, weighted arithmetic mean, etc. Some commonly used functions are given in Table 10.1 below.

Minimum	$E = min(e_1, e_2, \ldots, e_n)$	$r = -\infty$
Harmonic mean	$E = 1/(w_1/e_1 + w_2/e_2 + \ldots + w_n/e_n)$	$r = -1$
Geometric mean	$E = e_1^{w_1} \cdot e_2^{w_2} \ldots$	$r = 0$
Weighted arithmetic mean	$E = w_1e_1 + w_2e_2 + \ldots + w_ne_n$	$r = 1$
Square mean	$E = \sqrt{w_1e_1^2 + w_2e_2^2 + \ldots}$	$r = 2$
Maximum	$E = max(e_1, e_2, \ldots, e_n)$	$r = +\infty$

Table 10.1. Aggregate Function Spectrum.

The aggregation functions represent different degrees of conjunction and disjunction of negotiation data conditions. They can be selected by a user to suit different decision situations and for the selection of different products and services. For example, the maximum aggregation function is suitable when one or more of the negotiation conditions are acceptable to the user. In this case, the maximal value among all the preference scores derived for the negotiation conditions will be used as the global preference score. In another example, if CPU speed is most important to a client and s/he is 90% satisfied (i.e., a preference score of 0.9) with the speed of the computer under consideration, the preference scores of the rest of the attributes can be ignored. In this case, the global preference score is 0.9. At the other end of the spectrum, the Minimum function would use the minimal score among all the preference scores as the global preference. In the above example, if a client is only 10% satisfied with the speed of the CPU, the global score is 0.1 even though s/he may be totally satisfied with all other attribute values. As pointed out above, the aggregation functions defined by a decision maker represent different degrees of conjunction and disjunction. A naïve user, who cannot be expected to know the mathematics behind these functions, can be asked to select a value from the range (0,1) to express his/her desired degree of satisfaction and the system can map the value to the proper aggregation function.

In this work, we adapt the decision model described above and implement a Cost-benefit Evaluation and Selection (CBES) server (Yu (2001); Liu, Yu, Su, and Lam (2002)) based on the above model. Forms accessible through Web browsers are provided for companies to register their preference scoring and aggregation methods and the costs associated with different features of a product or service. When CBES is presented with a structure of values for describing a product or service, it uses the registered preference scores, aggregation method, and cost information to derive a global cost-benefit measure for the product/service. CBES and its services are useful for supporting decision-making in supplier selection, negotiation proposal evaluation, and evaluation of responses to a request for quote. They can be used to perform cost-benefit evaluations of all types of business specifications that can be defined by a hierarchical structure of attributes and values.

5.4 Negotiation Technology and Server

The services provided by the ETR, CSP and CBES servers were used to implement a Negotiation Server, which has been reported in Huang (2000); Hammer et al. (2000); Su, Huang, and Hammer (2000); Su et al. (2001), and Li (2001). Negotiation servers are replicated and installed at the sites of business organizations that conduct negotiations.

Each negotiation server would carry out a negotiation process on behalf of its client, based on some registered bargaining strategies (managed by an ETR server), product/service constraints (managed by a CSP server) and costs, preference scoring methods, and aggregation functions (managed by the CBES server). Negotiation proposals and counter-proposals are passed between a pair of negotiation servers in a bilateral bargaining situation. The negotiation server uses CSP to match the received proposal against the receiver's produce/service specification. If any violation of constraint is detected, it will post an event that corresponds to the violation to trigger some concession rule to relax the receiver's constraint and generate a counter-proposal to its counterpart. If no constraint violation is found in the negotiation proposal and the proposal specified several value combinations that are acceptable to the sender of the proposal, the CBES server is used to evaluate these value combinations to determine the best cost-benefit combination to present to the receiver for his/her acceptance.

6. Summary and Conclusion

The focus of our R&D effort has been to: (1) research a number of enabling technologies to support distributed applications in a supply chain, and (2) develop these technologies as collaborative e-business servers and integrate them to form a scalable, information infrastructure to enable e-supply chain management. In the introduction section, we have identified a number of requirements for an e-supply chain management system, which include active distributed objects, business event and rule management, quote evaluation and selection, automated negotiation, cost-benefit evaluation and selection, and order fulfillment. To satisfy these requirements, we suggest that a number of core technologies and their implementations in the form of servers are needed to provide various e-services to different business entities in a supply chain.

In this chapter, we have presented a general model of ESCM along with a representative scenario in Section 2. In Section 3, we presented an active object model for modeling inter-enterprise resources as active distributed objects. We presented the architecture of a distributed supply chain management system, which consists of a network of ESCM nodes, each of which contains a set of integrated servers to provide various e-services. Some core technologies and their implementations as servers are described in Section 5. ESCM offers a number of desirable features. First and foremost is the flexibility offered to business entities in defining their own rules according to their own business policies and strategies. Hence there is no need for business entities to hard code their business rules in application systems. Second, since the rules that control the business activities are installed and processed by the ETR servers, which are installed at various business entities' individual sites, an entity's privacy and security are safeguarded. Third, ESCM's event, event filtering, and event notification mechanisms keep business entities in a supply chain better informed, provide more timely information about business events, and allow information sharing through the events. Fourth, our system allows business processes to adapt to the runtime behavior of the system, allowing more flexibility in business operations. Fifth, our system can be used to manage a large supply chain by replicating ESCM nodes at a large number of Web sites.

We conclude this chapter by addressing the scalability issue, which is important in any information infrastructure development. "Scalability" defined at the system level is different from the inter-enterprise level. At the system level, scalability means that a system is able to sustain the performance when the data size, the number of users, the number of jobs, etc., continue to increase. At the inter-enterprise level, it has the additional requirement that the information infrastructure must accommodate different types and growing numbers of participating enterprises and enterprise resources without compromising *reliability, adaptability, flexibility* and *availability*. The architecture design and approaches taken to implement these e-services ensure that the other requirements at the inter-enterprise level will also be met. For example, the replication of the servers, and thus their e-services at multiple nodes, will ensure the high availability of these services. The support of parameterized events, event filters, dynamic rules, and parameterized rules ensures their adaptability to dynamic conditions of business. The integration of event and rule services with business process modeling and control to achieve active process management ensures flexibility in modifying inter-enterprise business processes. The registration services provided by the Negotiation Server and the Cost-Benefit Evaluation Server will allow new companies and users to easily specify their interests and requirements when they participate in the e-business enterprise.

References

Chakravarthy, S., Anwar, E., Maugis, L., and Mishra, D., 1994. "Design of Sentinel: An Object-Oriented DBMS with Event-based Rules," Information and Software Technology 39(9), London.

Dayal, U., 1998. "The HiPAC Project: Combining Active Databases and Timing Constraints," ACM SIGMOD Record 17(1).

Dujmovic, J.J., 1975. Extended continuous logic and the theory of complex criteria, Journal of Belgrade, Series on Mathematics and Physics 537, 197- 216.

Hammer, J., Huang, C., Huang, Y.H.,Pluempitiwiriyawej, C., Lee, M., Li, H., Wang, L., Liu, Y., and Su, S.Y.W., 2000. "The IDEAL Approach to Internet-Based Negotiation for E-Commerce," Proceedings of the International Conference on Data Engineering, SanDiego, CA, Feb. 28 - March 3, 2000.

Huang, C., 2000. "A Web-based Negotiation Server for Supporting Electronic Commerce", Ph.D. Dissertation, Department of Computer and Information Science and Engineering, University of Florida, 2000.

Lam, H. and Su, S.Y.W., 1998. "Component Interoperability in a Virtual Enterprise Using Events/Triggers/Rules," Proc. of OOPSLA '98 Workshop on Objects, Components, and Virtual Enterprise, Vancouver, BC, Canada, Oct. 18-22, 1998, pp. 47-53.

Lee, H.L., Padmanabhan, V. and Whang, S., 1997. "Information Distortion in a Supply Chain: The Bullwhip Effect." Management Science 43(4), 546-558.

Lee, M., 2000. "Event and Rule Services for Achieving a Web-based Knowledge Network," PhD Dissertation, Department of Computer and Information Science and Engineering, University of Florida, 2000.

Lee, M., Su, S.Y.W., and Lam, H., 2001. "Event and Rule Services for Achieving a Web-based Knowledge Network," Proceedings of the First Asia-Pacific Conference on Web Intelligence (WI-2001), Maebashi City, Japan, Oct. 23-26, 2001.

Li, H., 2000. "Automated E-Business Negotiation: Model, Life Cycle, and System Architecture", Ph.D. Dissertation, Department of Computer and Information Science and Engineering, University of Florida, 2000.

Liu, Y., Yu, F., Su, S.Y.W., and Lam, H., 2002. "A Cost-Benefit Evaluation Server for Decision Support in E-Business," accepted for publication in the Journal of Decision Support Systems and Electronic Commerce, 2002.

Lodha, R., 2002. "An event-Trigger-Rule based Supply Chain Management System over the Internet", Master Thesis, Department of Computer and Information Science and Engineering, University of Florida, 2002.

Parui, U., 1999. "Knowledge Profile Manager for Supporting Event-trigger-rule Services on the Internet," Master's Thesis, Department of Computer and Information Science and Engineering, University of Florida, 1999.

Shenoy, A., 2001. "Persistent Object Manager," Master's Thesis, Department of Computer and Information Science and Engineering, University of Florida, 2001.

Stonebraker, M., Hanson, E.N. and Potamianos, S., 1988. "The POSTGRES Rule Manager," IEEE Transactions on Software Engineering 14(7).

Su, S.Y.W., Dujmovic, J., Batory, D.S., Navathe, S.B., and Elnicki, R., "A Cost- Benefit Decision Model: Analysis, Comparison, and Selection of Database Management Systems." ACM Transactions on Database Systems 12(3).

Su, S.Y.W., Huang, C. and Hammer, J., 2000. "A Replicable Web-based Negotiation Server for E Commerce," Thirty-third Hawaii International Conference on System Science (HICSS-33), Wailea, Maui, Hawaii, January 4-7, 2000.

Su, S.Y.W., Huang, C., Hammer, J., Huang, Y., Li, H., Wang, L., Liu, Y., Pluempitiwiriyawej, C., Lee, M., and Lam, H., 2001. "An Internet-based Negotiation Server for E-Commerce", VLDB Journal 10(1).

Su, S.Y.W. and Lam, H., 2000. "IKnet: Scalable Infrastructure for Achieving Internet-based Knowledge Network," Proceedings of the International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet, l'Aquila, Rome, Italy, July 31-Aug. 6, 2000.

Su, S.Y.W. and Yu, T.F., 1997. "Distributed Information Mediation and Query Processing in a CORBA Environment," International Symposium on Digital Media Information Base, Nara, Japan, Nov. 26-28, 1997.

Widom, J., 1996. "Active Database Systems: Triggers and Rules for Advanced Database Processing," Morgan Kaufmann, San Francisco, CA, 1996.

Yu, F., 2001. "A Cost Benefit Evaluation Server for supporting general decision-making," Master's Thesis, Department of Computer and Information Science and Engineering, University of Florida, 2001.

III

FROM RESEARCH TO PRACTICE

This page intentionally left blank

Chapter 11

THE STATE OF PRACTICE IN SUPPLY-CHAIN MANAGEMENT: A RESEARCH PERSPECTIVE

Leroy B. Schwarz
Krannert Graduate School of Management
Purdue University
West Lafayette, Indiana 47907
lschwarz@mgmt.purdue.edu

Abstract In this chapter, I will describe examples of state-of-the-art practice in supply-chain management; e.g., vendor-managed inventory, quick response, and other contemporary systems, such as Wal-Mart's *RetailLink.* The perspective will be that of what I call the *IDIB* Portfolio; i.e., what *Information (I), Decision-Making (D), Implementation (I),* and *Buffer (B)* systems are employed in managing real-world supply chains. Most operations-research models consider only two components of this portfolio: the decision-making and the buffer systems. More specifically, most operations-research models involve selecting a decision-rule to minimize expected buffer (i.e., inventory-holding) and backorder cost given a fixed level of information. Implementation as a decision variable is typically ignored. However, in the real world, with changing information, communication, and implementation technologies, supply-chain management can – and *should* – be viewed as changing the nature of the *entire* IDIB Portfolio. After interpreting current practice from the perspective of the IDIB Portfolio, I will forecast *future* practice using Collaborative Planning, Forecasting, and Replenishment (CPFR) as an example. I will describe the elements of CPFR, identify companies that are using it, and the challenges they face in realizing its potential. Finally, I will identify research opportunities in CPFR, and, more generally, research opportunities involving the IDIB Portfolio.

1. Introduction

We've all seen one version of it or another: The consulting company's Power-Point slide representing the evolution of supply-chain management (SCM). A staircase ascends left to right: on the bottom step is "Basic Supply-Chain Management", typically intra-company MRP or ERP systems; and on the top step, "Advanced Supply-Chain Management", typically described as "wireless, broad-band, web-based, truly collaborative, etc.". The next slide asks: *"Where is your company?"*

Sridhar Tayur (Carnegie Mellon University) displayed one of these staircase slides at a recent meeting of supply-chain thought leaders at Harvard University. Marshall Fisher (University of Pennsylvania) asked: "Does anyone know of any companies that are at or near the top step?" No one raised a hand. That slide, Marshall's question, and *no one's* answer provided the motivation for this chapter.

In what follows I will do three things: First, I will describe the state-of-the-art in supply-chain management practice. A caveat: my assessment will be general; no details of specific buyer-supplier practice will be described. Second, I will introduce a paradigm – called the "IDIB Portfolio" – for understanding the evolution of supply-chain management to date, and for predicting its future. Third, using this framework, I will suggest topics for research.

My focus will be on managing the link(s) between buyers and suppliers that are independently owned and managed. Although centrally-owned and managed links provide a valuable benchmark, the real challenges in supply-chain management involve two or more independently-owned and managed companies. Two fundamental challenges are posed in improving the management of such supply chains: (1) the development of techniques to improve overall supply-chain performance (e.g., increasing total supply-chain profit); and (2) the development of contracting mechanisms that will motivate *all* the partners to implement these techniques. In other words: (1) how to "enlarge the pie"; and (2) how to "provide larger slices" to all the partners. I will focus on the first challenge – the development of techniques to improve overall performance. See Cachon, 2004 for a review of the supply-chain literature on the management of incentive conflicts with contracts.

This chapter is organized as follows: Section 2 provides an overview of contemporary supply-chain management systems. In Sections 3-4 I will introduce the IDIB Portfolio paradigm, describe what "managing the IDIB Portfolio" means, and contrast the IDIB Portfolio paradigm with the operations-research paradigm. Section 5 uses the IDIB Portfolio to provide a perspective on contemporary supply-chain management practice. Section 6 describes two "Axioms of the IDIB Portfolio". In section 7, I use these axioms to forecast the future of supply-chain management. Section 8 provides an overview of Collaborative Planning, Forecasting, and Replenishment (CPFR), which, I believe, is one "contender" for the future of supply-chain management. In Section 9, I identify several research topics in CPFR, in Section 110, several research topics in supply-chain collaboration, and, in Section 11, several research topics related to the IDIB Portfolio paradigm. Section 12 cites closely-related references. Section 13 provides a summary.

2. An Overview of Contemporary Supply-Chain Management Systems

Vendor-Managed Inventory (VMI), introduced by Kurt Solomon Associates in 1992 (`http://www.kurtsalmon.com`), is perhaps the most widely-known system for managing supply chains. Under VMI, the buyer authorizes the supplier (i.e., vendor) to manage the inventory of a set of stock-keeping units (SKUs) at the buyer's site(s) under agreed-upon parameters (e.g., minimum and maximum inventory targets). The buyer provides the supplier with sales and/or inventory-status information; and the supplier makes and implements decisions about replenishment quantities and timings.

VMI reduces information distortion, which is one cause of the "bull-whip" effect (Lee et al., 1997). In addition, VMI provides the supplier with the opportunity to better manage its own production, inventory, and transportation costs. (See, for example, Çetinkaya and Lee, 2000). In exchange, the buyer typically receives price discounts or improved terms of payment from the supplier.

Quick Response (QR) was innovated by Milliken & Company (`http://www.milliken.com`) in the early 1990's and subsequently codified by the Voluntary Interindustry Commerce Standards (VICS) Association. QR has four levels of application and technology. Levels 1 and 2, for example, involve retailer inventory-status information-sharing and automatic order-processing between retailer and supplier. Levels 3 and 4 include VMI and cross-docking warehouses. See Fiorito et al., 1994 for more information.

Although VMI and QR might be the "best-known" management systems among both practitioners and academics, perhaps the most *highly-regarded* systems are proprietary systems developed by large retailers, such as Wal-Mart's *RetailLink,* Kmart's *Workbench,* and Target's *Partners Online.* Although the detailed inner workings of these systems are closely-guarded secrets, they all have two common characteristics: (1) the sharing of transactions-level data among partners; and (2) the use of agreed-upon metrics (e.g., in-stock, inventory-turnover, and on-time delivery measures) and targets to assess partner performance. *RetailLink,* for example, captures sales, inventory, and delivery-related data for every SKU at every Wal-Mart facility (i.e., store and distribution center) and uploads it to a central database at least every 24 hours. These data, and metrics based upon them, are made available to *every* manager and *every* company up or down the supply chain whose performance is related to this SKU.

How this shared information is used – in particular, whether the decisions based on it are made centrally or decentrally – depends on the specific partnership agreement, and the type of product(s) being managed. Similarly, who does the implementation and how it is done is specific to the partnership and product(s). To illustrate: Wal-Mart generally centralizes decision-making but decentralizes implementation for Wal-Mart and Sam's Club facilities, while delegating decision-making and implementation to its suppliers. However, regardless of who makes or implements the decisions, the quality of the decisions and their implementation are continually monitored by all parties and compared to the agreed-upon targets.

Before taking a closer look at current supply-chain practice, I will introduce a simple paradigm that will be useful in interpreting it.

3. Information, Decision, Implementation, and Buffer (IDIB) Systems

Managing anything, including managing a supply chain, can be viewed as four related activities: (1) *getting information* for decision-making; (2) *decision-making;* (3) *implementing decisions;* and (4) *buffering* against imperfections in (1)-(3). Correspondingly, every organization has systems for performing these four activities.

The Information System

The role of the "information system" is to provide past, present, and future-oriented information for decision-making. This information might be about demand, costs, materials, capacities, etc. An *ideal* Enterprise Resource Planning (ERP) system should be capable of capturing and providing *all* this information. In practice, however, information "in the information system" is captured and stored in dozens of different ways, among them paper records and computer files. In many organizations, much of this information is, literally, in the heads of management personnel.

Hence, a firm's information system is the less "a thing" and more a collection of "things": the collection of all business processes – formal and informal – plus all the information technologies and systems that provide information for decision-making. In this sense, information systems are "decision-support systems". However, I prefer the label "information system" since most decision-support systems are limited in their domain, and, hence, are a subset of the information system.

The overall quality of an information system's information about anything can be measured as some combination of: (1) its *accuracy;* that is, the correspondence between past, present, and future reality and what the information system reports (or reported) about it; (2) its *leadtime;* i.e., the time between an event and when the information system reports it; (3) its *level of aggregation;* i.e., the categories and units in which information is provided; and (4) its *horizon;* i.e., how far into the future (or past) the information system looks.

At their worst, information systems provide grossly inaccurate, even irrelevant, information. "Better" information systems typically provide some historical information and some current-status information. Still better information systems provide future-oriented information, such as demand forecasts and cost projections. A "perfect" information system would be the proverbial crystal ball, capable of seeing perfectly into the past, present, or future.

Another important characteristic of a company's information system is its cost; that is, the cost of the people, equipment, facilities, and processes that, together, comprise the information system. Typically, for a given technology, the cost of an information system is an increasing and marginally-increasing function of its overall quality. In other words, improving a given information system costs more, and each additional increment in quality costs more than the last.

From a management perspective, of course, the *value of an information system* doesn't depend on the *quality of the information* it provides, but on the *quality of the decisions made* based on this information.

The Decision-Making System

The role of a decision-making system, of course, is to make decisions using the information provided by the information system. Decision-making takes many different forms and is performed by many different individuals or groups. Decision-making occurs throughout every organization, from the shop-floor to the executive suite. Strategic decisions (e.g., the organization of the supply-chain, product offerings) are typically made at the executive level. Managerial decisions (e.g., the master schedule, order-promising) are typically made by middle managers. Tactical decisions (e.g., processor assignment, workload sequencing) are typically delegated to the shop floor. The decision-making processes themselves might be informal, even intuitive (e.g., "Ned uses his 25 years of experience to assign workloads to processors"), or they might be codified, even regulated. For example, in US pharmaceutical manufacturing, most of the production decisions (e.g., the process steps and lot sizes) have been specified and approved by the FDA and, hence, *must* be adhered to. The corresponding decision rules might be simple – for example, in master-scheduling "run-out rules" are often used – or mathematically sophisticated (e.g., master-schedule optimization using mathematical programming). The decision-rules might be formal and automated; or, they might be totally within the head of some individual or group of individuals. Typically, important decisions are the result of a complex set of activities, some logical or based on management judgment, some simply guess-work.

The overall quality of a decision-making system's decisions can be measured as some combination of: (1) the *optimality* of the decisions made; and (2) the decision-making *leadtime.* "Optimality" means how desirable the decision is – given the quality of information provided by the information system – with respect to cost, profit, or some other measure of utility. The "decision-making leadtime" is the amount of time it takes to make the decision once the appropriate information has been provided. This leadtime might be short or long; it might be a fixed amount of time, or variable; it might be known or unknown in advance.

Like information systems, the cost-drivers for decision-making systems are the people, equipment, facilities, and processes that, together, comprise it. And, like information systems, for a given technology, the cost of a decision-making system is typically an increasing, and marginally-increasing function of its overall quality. In other words, improving a given decision-making system costs more, and each additional increment in quality – for example, each increment in the desirability of its decisions or each decrease in leadtime – costs more than the last.

Finally, like information systems, the *value* of a decision-making system doesn't depend only on its quality characteristics. In particular, a "perfect" decision that doesn't get implemented is of *no* value.

The Implementation System

Implementation usually involves some "paperwork" to authorize or initiate activity. For example, a decision to ship 200 units from Dock 4 at 12 o'clock on April 5th will typically involve inventory-withdrawal authorizations, transportation requisitions, etc. Often, some preliminary actions must also be taken. For example, if 200 units aren't in inventory, then production decisions and their corresponding implementations must take place.

The overall quality of an implementation system can be measured as some combination of: (1) the *implementation leadtime;* and (2) *imple-*

mentation accuracy. The "implementation leadtime" is the amount of time required to make the decision happen; in other words, the time between making the decision and having the corresponding actions completed (i.e., implemented). For example, the amount of time it actually takes to ship the 200 units from Dock 4 once the decision has been made to do it. Like decision-making leadtimes, implementation leadtimes might be short or long, fixed or variable, known or unknown in advance. "Implementation accuracy" measures how closely the implementation matches the decision. Perfect accuracy means that the implementation perfectly matched the decision; for example, that exactly 200 units were shipped from Dock 4 at 12 o'clock on April 5th. In practice, implementation is seldom perfect: differences may be small, as in a tightly-controlled JIT system, or large, as in high-density chip fabrication, where yield losses are unpredictable and difficult to control.

Often trade-offs occur between the implementation leadtime and the accuracy-of-implementation. The phrase "quick-and-dirty", for example, means that decision-making and implementation are "quick" (i.e., that their combined leadtime is short) but that the decision and/or its implementation are "dirty" (i.e., that the decision isn't very desirable and/or the implementation isn't accurate).[1]

Like information and decision-making systems, implementation systems cost money; that is, the cost of the people, equipment, facilities, and processes that involve making the decision happen. And, like information and decision-making systems, for a given level of technology, the cost of an implementation system is typically an increasing and marginally-increasing function of its overall quality. In other words, improving a given implementation system – that is, making it faster or more accurate – costs more, and each additional increment in quality costs more than the last.

The Buffer System

In a perfect world, information systems would provide perfect information and decision-making systems would make perfect decisions. Implementation would be perfect, too. However, in the real world, *none* of these three systems is *ever* perfect. Management systems compensate for these imperfections using buffers and buffer systems.

What are buffers? Unlike information, decision-making, and implementation systems, which can be realized in a virtual infinity of different

[1] In fact, decision-making and implementation are often iterative. For example, a decision might be made; then, during the process of implementation, new information is revealed that might lead to modifying the decision, etc.

forms, buffers come in *only* three basic forms: inventory, leadtime, and capacity. A "buffer system" is a combination of inventory, leadtime, and capacity buffers, in various amounts, located within what might be called the "management-system supply chain". Often, for example, inflated leadtimes or extra capacity are imbedded in the information, decision-making, and implementation systems. Work-in-process inventories are typically found throughout the implementation system, with raw-material inventories at the beginning and finished goods at the end of the implementation chain. Leadtime buffers (e.g., inflated promise dates) are typically found at the interface between the points of delivery and the customers.

Re-thinking Buffers

The best way to understand – indeed, appreciate – the roles that buffers play in a management system is this: *forget everything you've thought about them until now.* In particular, put aside any notion that buffers are inherently "bad".

Yes, buffers are often thought to be "bad". Why? Perhaps it's because buffers cost money. Yet, the other three components of a management system cost money, too. It is also worth noting – and peculiar, I think – that despite the fact that information, decision-making, and implementation systems *also* cost money, these three elements of a management system are generally thought to be "good". Further, it is generally accepted that "improving" – that is, increasing the capabilities of – an information, decision-making, or implementation system is a "good thing". On the other hand, it is generally thought that "improving a buffer" *must* mean reducing its capabilities or eliminating it. *Put this notion aside, too.*

Another reason why buffer systems might be thought to be "bad" is that whatever amount of buffering management chooses to provide, it is typically the *wrong* amount. In other words, the amount of buffering is either too much – for example, leftover inventory at the end of a selling season – or not enough – for example, that sales were lost despite the fact that "lots" of safety stock was provided.

I believe it is more useful – and accurate – to think about buffers in terms of the role they play in a management system: to compensate for imperfections in the information, decision-making, and implementation systems. From this perspective, I believe that buffers can only rightfully be called "too big" if they use *more* resources than necessary to do what they are intended to do: to compensate for imperfections *elsewhere* in the management system. Similarly, buffers should only be thought of as "bad" if they *imperfectly* compensate for those imperfections.

What would a "perfect" amount of buffering be?

Consider a business scenario like that of the newsvendor model, and a management system whose information system, *I*, perfectly forecast customer demand. Assume that this perfect forecast is provided to decision-making, *D,* and implementation, *I*, systems that are capable of providing exactly this number of units at the instant they are demanded in precisely the right quantity. How much buffering, *B,* is required to perfectly satisfy customer demand? None. In other words, the "perfect" amount of buffering in *this* system *is* zero.

Now change the management system slightly so that the information system can only provide a probability distribution of future demand. Given this level of imperfection in the information system, is *any* management system capable of *always* satisfying customer demand *without* lost sales or unused inventory, capacity, or leadtime?

The answer is "yes", *provided* that the decision-making and implementation systems are perfect; that is, provided that, once demand is known, management is capable of deciding to provide this amount, and provided that the implementation system is capable of producing precisely this amount *instantaneously.* Here, too, the perfect amount of buffering would be zero. On the other hand, *unless* the decision-making and implementation systems are perfect, *some* buffering will be required. How much and what kind of buffering?

Suppose that the "IDI_ systems" – that is, the combined information, decision-making, and implementation systems – are perfect with respect to leadtime, but imperfect with respect to quantity. In other words, *some* amount will be provided at exactly the right time to satisfy demand, but that the quantity won't necessarily equal the quantity demanded. In order to avoid lost sales, *inventory buffering* will be required, the amount depending on the overall imperfection, or variance, in the quantity the IDI_ provides. As the variance of this imperfection increases, in order to compensate, the corresponding amount of buffer inventory must also increase. In the extreme, as this variance increases to infinity, the "perfect" amount of buffering also increases to infinity.

Next, suppose that IDI_ systems above are *im*perfect in quantity *and* leadtime. In order to satisfy demand at the time that it occurs in the quantity demanded, a *leadtime buffer* must be added to the inventory buffer. And, as the uncertainty in "supply leadtime" increases, the corresponding leadtime buffer must also increase.

Finally, since in many situations, the uncertainty arising from the IDI_ systems is a consequence of scheduling conflicts on some constrained resource(s), capacity buffers can usually be substituted for some of the inventory and leadtime buffers. Hence, the appropriate *forms* of buffering – inventory, leadtime, or capacity – depend on the *nature* of the imperfection(s) in the IDI_ systems.

Finally, then, given some form of buffering, what is the "perfect" amount?

The "Perfect" Amount of Buffering?

If the role of the *management system* – the *IDIB* – is to provide what is required or demanded at the time and in the quantity that's required or demanded; and if the role of a buffer is to compensate for imperfections in the IDI_ systems, then the "perfect" amount of leadtime, capacity, or inventory is the amount that does just that and no more: the amount that provides precisely what's required or demanded despite the imperfections in the IDI_ systems. Hence, the "perfect" amount of buffering is *never* zero unless the IDI_ systems are capable of perfection. Furthermore, as the combined imperfections in the IDI_ increases, the corresponding perfect amount of buffering must also increase.

> An aside on the "Zero Inventory" concept often associated with Just-in-Time systems: The concept of zero inventory, or, more generally, zero buffering can be extremely important in identifying the nature and magnitude of imperfections in an IDI_ system. In other words, reduce the amount of buffering and "see" what imperfections are revealed. Often such experiments uncover imperfections in the IDI_ systems that are inexpensive to reduce or eliminate. If so, then this should be done and the corresponding amount of buffering should be permanently reduced. However, to the extent that imperfections in the IDI_ systems remain, the perfect amount of buffering isn't zero.

Hence, the *perfect* amount of buffering depends on the *imperfections* in the combined IDI_ systems. If the amount of imperfections – that is, the uncertainties associated with the IDI_ systems – increases, then the overall level of buffering also must be "improved" – increased – in order to compensate.

The Cost of Buffering

Like all the other components of a management systems, the cost of a buffer system is typically an increasing, and marginally-increasing function of its overall quality. In other words, improving – that is, increasing – the capability of any given buffer costs more, and, typically, each additional increment in quality costs more than the last.

An Alternative Perspective: the "Optimal" Amount of Buffering

The operations research (OR) paradigm provides an alternative perspective on buffers, and, based on this perspective, defines the "optimal" amount of buffering. The perspective of the OR paradigm is as follows: The realized amount of buffering provided by a given management system is the difference between what reality requires or demands and what the management system provides. For example, if demand was 100 units and the management system provided 120 units, then the buffer was 20 units. Or, if management provided capacity of 40 hours and capacity of 50 hours was required, then the buffer was –10 units. A negative buffer, whether it's inventory, capacity, or leadtime, means that the management system didn't provide enough of what was required or demanded; a positive buffer means that too much was provided.

Correspondingly, the OR paradigm takes the view that the cost of the buffer system is the cost of all the positive buffering *plus* all the costs associated with negative buffering. In other words, the cost of the buffer system is the cost of all the *unused* inventory, capacity, and leadtime that the management system provided – measured *after* demand for that resource occurs – *plus* the cost of all the corresponding shortages of inventory, capacity, and leadtime. (e.g., lost sales, backorder, goodwill cost).

The newsvendor model is, perhaps, the best-known example of the OR paradigm. The newsvendor model chooses the optimal "target inventory" based on three "drivers": the per-unit cost of "not enough" buffering (i.e., opportunity cost of lost sales), the per-unit cost of "too-much buffering" (i.e., the out-of-pocket leftover cost), and the probability distribution of demand. The expected-cost minimizing target inventory is provided by the well-known "newsvendor fractile" of the cumulative demand distribution. The corresponding "safety-stock" (inventory) buffer is measured, *a priori,* as the difference between the chosen target inventory and the expected customer demand.

What's Wrong with the OR Paradigm's View of Buffers ?

There is nothing "wrong" with the OR paradigm or with its view of buffers. "Near-sighted" would be a better description.

In particular, the OR-paradigm's view of the too-much and not-enough costs associated with the imperfections of a management system is consistent with the IDIB-paradigm perspective. In other words, once, say, demand occurs, the relevant costs incurred because of an imperfect IDIB Portfolio are those associated with providing too much or not enough of whatever was demanded.

But why should these too-much and not-enough costs *only* be associated with – if you will, "blamed" on – the buffer system? Yes, buffers are typically imperfect, but what about the imperfections in the IDI_ systems? More on this below.

One way of looking at the difference between the OR and the IDIB paradigms is that the IDIB paradigm generalizes the OR paradigm. Specifically, the IDIB paradigm takes the view that the quality of *all* four components of a management system – not just the buffer system quality (i.e., size) – are decision variables; second, that the underlying quality-cost function for each of these components – that is, the cost to move them in the direction of perfection – is increasing and marginally increasing. Finally, that there is a cost associated with the entire IDIB Portfolio – again, not just the buffer system – for failing to provide whatever is required or demanded. Hence, the "optimal" IDIB Portfolio is the IDIB Portfolio that minimizes the total costs of *all* of its components *plus* the cost of failing to provide precisely what is/was required or demanded.

From this point of view, the OR paradigm is "near-sighted" to the extent that it takes the IDI_ as fixed, and focuses only on picking the amount of buffering that minimizes the corresponding too-much and not-enough buffer cost.

We will address the question of the optimal IDIB in Section 11. We turn now to the concept of the "IDIB Portfolio".

4. The IDIB Portfolio

I label the *combination* of these four components of a management system – the information, decision-making, implementation, and buffer systems – a "portfolio", because, like a financial portfolio, each of these systems involves an investment of dollars. And, like the performance of an investment portfolio, the performance of a management system depends on how well its components perform in *combination,* not as separate components. Finally, in assembling an IDIB Portfolio, as in assembling a financial portfolio, an almost unlimited number of combinations can be chosen. As an illustration, suppose the goal is to manage a supply chain to provide a 95% customer fill-rate at the retail level. This might be provided by managing every link of the chain with low-quality information, decision-making, and implementation systems, but large inventory buffers everywhere. Or, without changing the management of the other links in this supply chain, the manufacturer might substitute buffer capacity for some of its finished-goods inventory and still offer the same service to the distributor. Similarly, the distributor might choose higher cost of express delivery (from the manufacturer) in order to reduce the expediting and inventory-holding cost on its safety-stock (buffer) inventory. Three different IDIB Portfolios all providing the same level of customer service.

Which is best?

If the goal of a management system is to maximize profit, then the "best" IDIB Portfolio is the portfolio with the least total cost: that is, the total cost of all the people, facilities, equipment, and technology associated with providing information, making decisions, implementing them, and buffering to compensate for imperfections in the IDI_ systems, *plus* the cost of lost sales, expediting, and goodwill loss resulting from failing to do all this perfectly.

Managing the IDIB Portfolio

"Managing" the IDIB Portfolio means making decisions about the nature and quality of its four components: the information, decision-making, implementation, and buffer systems. The *ideal* is to select the quality of each component, plus the cost of lost sales, etc., so that total cost is minimized. Is "managing the IDIB Portfolio" amenable to the tools of operations research? I think it *could* be, but most OR models are much too myopic to be called "IDIB optimizers", much less, "IDIB improvers".

Consider the well-known newsvendor inventory model. In this model, the information system provides the probability distribution of customer demand and estimates of the costs associated with buying and selling newspapers. Attention is focused on the decision about the number of newspapers to have on hand at the beginning of the day in order to maximize the newsvendor's expected profit. The optimal decision-rule is well known: set this inventory equal to the "critical fractile" of the probability distribution of customer demand. The newsvendor's implementation leadtime is not explicitly considered. Instead, it is usually assumed that whatever this leadtime is, it is short enough so that, once the newsvendor has decided how much to order, the chosen quantity will be delivered on time. The basic newsvendor model also ignores implementation accuracy; that is, it is implicitly assumed that whatever quantity is ordered will be delivered. Extensions of the newsvendor model consider accuracy; that is, the correspondence between what is ordered and what is delivered (See Karlin, 1958 and Ehrhardt and Taube, 1987, for example; see Yano and Lee, 1995 for other references).

I believe the newsvendor model is representative of virtually every operations-research model of supply-chain management. That is, the quality of the information provided by the information system is *assumed* to be fixed. The costs associated with the information, decision-making are typically *ignored.* The costs associated with implementation are sometimes represented, but in a highly-stylized manner. For example, the only implementation cost associated with the newsvendor model is the marginal purchasing cost. Similarly, the only implementation cost associated with the EOQ model is the assumed-to-be-fixed order cost. The goal, as described above, is to determine the decision-rule that minimizes buffer-system cost. In IDIB Portfolio terms, the newsvendor model – like most supply-chain models – selects the decision-rule for the decision-making system ("D") that minimizes expected total buffer-system ("B") cost for a fixed quality of information, "I". The implementation "I" is represented either as a cost or as a constraint; not as a decision-variable.

"Managing the Newsvendor's IDIB Portfolio" is much more complex. It *does* involve selecting decision rules, but the objective is to minimize *total* portfolio cost, not just the cost of the buffer system. In particular, managing the newsvendor's IDIB Portfolio *also* involves assessing the cost and value of an information system that would provide more (or, possibly, even less!) precise information about customer demand. It *also* involves assessing the cost and value of implementation systems with different leadtimes and accuracies. More broadly, managing the newsvendor's IDIB portfolio might involve fundamental changes in the newsvendor's operations.

Suppose, for example, that the newsvendor was able to make and implement her/his ordering decisions any number of times during the day and receive those newspapers instantaneously (i.e., decision and implementation leadtimes are zero). In such a scenario, the newsvendor wouldn't inventory *any* newspapers, nor would he/she need a probability distribution of demand. Instead, the newsvendor would wait until a customer requested a newspaper and then provide it upon demand. Sound far fetched? Consider this:

For decades, the copier division of Xerox struggled with managing the inventory of owner's manuals for its copiers: how much inventory of which manuals to have on hand, and when to replenish this inventory. Xerox eliminated this problem by developing a system for printing and binding manuals upon demand; i.e., whenever assembly of a copier is scheduled, the printing of its manuals is also scheduled.

How was this done? By developing a system for implementing the decision to produce manuals whose leadtime is less than or equal to the time required to implement the decision to assemble the copier.

"Managing" the IDIB Portfolio is no easy task. First, the different components of the portfolio are often difficult to identify. For example, managers, who are nominally decision-makers, also often play a role in the information system; line personnel, who are nominally implementers, have roles to play in information, decision-making, and buffering. So, it is often difficult to separate the components of a firm's IDIB Portfolio. Second, many of the costs associated with a firm's IDIB Portfolio are "overhead" or "indirect" costs, which makes them difficult to estimate. These and other difficulties make it virtually impossible to find the truly optimal IDIB Portfolio; that is, the portfolio of information, decision-making, implementation, and buffering whose combined cost *plus* the cost of failing to provide enough of whatever was demanded or required (e.g., backordering or expediting cost) is the minimum possible cost.

Nonetheless, it is often relatively easy to verify that one given IDIB Portfolio has lower total cost than another. In the Xerox example above, it was relatively easy to verify, using back-of-the-envelope estimates, that Xerox's "new" portfolio was an order of magnitude less expensive than its old one.

Xerox's old information system ignored the fact that *Xerox management* decided which copiers to produce and when to produce them. Instead, it assumed that demand for a given manual was provided by a probability distribution. Xerox's new IDIB Portfolio uses information that management has scheduled a given copier to be assembled, and then implements the decision to print its manuals in a short enough leadtime so that the manual can be packaged with the copier at the end of the assembly line.

I believe that experienced operations researchers, practitioners, and consultants already recognize the tradeoffs that the IDIB Portfolio makes explicit. For example, consultants will often prescribe a less-than-perfect, heuristic decision-rule because it is less demanding of the information system and/or easier to implement, particularly if these imperfections are relatively inexpensive to buffer against. On a broader level, sensitivity analysis, which is a well-founded tool of the operations research theorist, can be viewed assessing the impact of imperfections in the quality of the information and/or decision-making systems on system performance, and, hence, the level of buffering that might be required. For example, sensitivity analyses on the basic EOQ model can be viewed as assessing the sensitivity of lot-sizing decisions to inaccuracies (i.e., imperfections) in the information required to support it (e.g., estimates of company inventory-holding cost or set-ups). See Lowe and Schwarz, 1983, for example.

Similarly, but at a metaphysical level, operations-research theorists often prefer a less realistic (i.e., less perfect) model to a more realistic model because of the insight its analysis provides.

From this point of view, the fundamental difference between the IDIB paradigm and the OR paradigm is that experienced operations researchers, practitioners, and consultants make these tradeoff *a priori;* that is, without explicitly identifying all the alternative levels of quality in information, decision-making, and implementation systems that might be chosen. Although this is understandable, such *a priori* choices necessarily lead to locally, not globally, optimal choices for the corresponding IDIB Portfolio.

The Role of Information Technology and Economics

What "happened" at Xerox that led to its development of a new IDIB Portfolio for managing manuals? Was it management's "discovery" that *their own decisions* created the demand for manuals? Not likely.

Obviously, the availability of technology played a role: printing-and-binding technology that could "quickly" produce an owner's manual. In IDIB Portfolio terms, technology whose implementation leadtime was shorter than the implementation leadtime to assemble a copier. The other element, of course, is economics: Given that technology *can* facilitate a "new" IDIB Portfolio, its adoption only makes sense if the total cost of the "new" portfolio is less than the total cost of the "old" portfolio. So, obviously, technology and economics play a substantial role in the development of "new" IDIB Portfolios.

Given the proven success of information technology (e.g., microcomputers and the internet) as a significant facilitator of improved information, decision-making, and implementation, I believe it is inevitable that information technology will continue to create the opportunities for "new" IDIB Portfolios. Further, to the extent that the cost of information technology continues to fall, following Moore's Law[2], these IDI_ systems will continue to become less and less expensive. And, buffers, "B"s – inventory, capacity, and leadtime – are, if anything, becoming more expensive over time.

Hence, I believe information technology and economics will continue to offer IDIB Portfolios to supply-chain managers whose total cost is less than today's[3]. The challenge to supply-chain managers is: Which

[2] The observation made in 1965 by Gordon Moore, co-founder of Intel, that the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented. Moore predicted that this trend would continue for the foreseeable future. In subsequent years, the pace slowed down a bit, but data density has doubled approximately every 18 months, and this is the current definition of Moore's Law, which Moore himself has blessed. Most experts, including Moore himself, expect Moore's Law to hold for at least another two decades. *Source:* `webopedia.com` (definition last modified in March, 1998).

[3] Although new technological capabilities are likely to be the primary drivers for changing an existing IDIB Portfolio, at least in the short term, it is important to note that even if technology were held constant, any significant change in the cost of one or more of the components of *any* given IDIB Portfolio is an equally-important driver. For example, holding everything else equal, a significant decrease in the cost of capital makes inventory and equipment buffering less expensive. Under these circumstances, increasing the quality of the buffer system and decreasing the quality of the corresponding information, decision-making or implementation system should reduce total cost.

new IDIB Portfolio to adopt, and when to adopt it? The challenge to supply-chain modelers is: What techniques and models will aid managers in choosing new IDIB Portfolios?

5. An IDIB Portfolio Perspective on Supply-Chain Management

Until fairly recently, virtually every link of every real-world supply chain was managed using a *very* crude IDIB Portfolio. Typically, the only information a buyer shared with its supplier was its current order; and the *most* information the supplier shared with the buyer was this order's planned or promised shipping date. Status information (e.g., order status, inventory status) and future-oriented information (e.g., planned orders and production) were seldom, if ever, shared, often because neither partner had easy access to its *own* information about them. Further, in those instances when one partner *did* have access to this information, the technology for sharing it either wasn't available or was very expensive. As a consequence, supply-chain "partners" – if we can call them that – were figuratively blind to one another. In the absence of useful supplier/customer information for decision-making, each "partner" made decisions that were focused on what little information was available, typically, internally-focused information, such as processor utilization, hot lists, etc. Consequently, the decision-making and implementation based on this low-quality information required *huge* buffers: large buffer inventories of raw materials, work-in-process, and finished goods *plus* buffer leadtimes and capacity.

Given the high cost of these buffers, it's no surprise that when low-cost technology for information-sharing between buyers and suppliers became available, innovators seized the opportunity to substitute low-cost information-sharing for these high-cost buffers, thereby achieved significant improvements in performance and/or reduced total cost.

Wal-Mart, of course, must be credited with introducing many of the technological innovations (e.g., bar-coding, satellite communication of point-of-sale information) associated with contemporary supply-chain management, and equally important, in *demonstrating* that substituting improved information for inventory and leadtime buffers reduced total cost.

And, so the "revolution" in supply-chain management began and continues today, the state-of-the-art represented by Wal-Mart's *RetailLink* system.

Yet, what is really different about these systems? From an IDIB perspective, these innovations are actually fairly modest, at least in terms of what *could* be innovated. Under Vendor-Managed Inventory (VMI), for example, the buyer delegates the making and implementation of inventory-replenishment decisions to the vendor. In order to do so, the buyer's information system provides the vendor with information about customer demand and inventory status. Levels 1 and 2 of Quick Response (QR) are the same, with the addition of automatic order-processing (i.e., implementation) between retailer and supplier. Levels 3 and 4 of QR include cross-docking warehouses (i.e., faster implementation of warehousing decisions). Even Wal-Mart's *RetailLink,* although awesome in its scale (e.g., in the level of detail provided, in the type and number of partners with access to it), is fundamentally: (1) a system for rapidly sharing transactions-level data and metrics about the past and the present with its suppliers; and (2) a centralized system for making and implementing decisions for its own facilities.

So, why the "significant improvements" from what I describe as "modest innovations"? There is an old saying that: *"In the land of the blind, the one-eyed man is king".* Historically, supply-chain "partners" were figuratively blind to one another. Add just a little vision – for example, under QR, customer-demand and retailer-inventory information – and, suddenly, there is no need for much of the buffering that had been required in the "land of the blind". Reducing inventory produces cash. Reducing capacity increases productivity, and, hence, profitability. Reducing leadtimes attracts more customers.

Given the huge payoffs that supply-chain partners have derived from sharing a modest amount of information or from delegating *some* decision-making and implementation from buyer to supplier, will more of the same yield even larger payoffs? Where will it end? Will it end?

The "Axioms of the IDIB Portfolio" suggest that long before information-sharing and/or delegation of decision-making and implementation becomes "total" between supply-chain partners – if it ever does – something even more revolutionary will happen.

6. The Axioms of the IDIB Portfolio

The 1st Axiom of the IDIB Portfolio is this: ***Given an existing IDIB Portfolio, increasing the quality of one of its components facilitates decreasing the quality of one or more of its other three components while maintaining the same level of customer service (e.g., fill-rate, leadtime).***

In an inventory-replenishment system, for example, reducing the leadtime to implement a replenishment decision, facilitates decreasing the safety-stock inventory or leadtime buffer – that is, *decreasing* its quality

– without affecting customer fill-rates, expected backorders, etc.. Or, in the same setting, decreasing the variance of the leadtime to *make* a replenishment decision by, say, one unit, facilitates increasing the corresponding *implementation* leadtime variance by one unit.[4] Many such tradeoffs are possible.

Schneider National Trucking Company's innovation of satellite tracking to locate its trucks allowed Schneider to reduce two of its costly buffers while *improving* customer service. Here are some details:

Before introducing its satellite-based information system for locating its thousands of trucks, Schneider dispatchers relied on periodic telephone calls from its drivers in order to learn where its trucks and drivers were. The corresponding uncertainty about where and when its own trucks would be available led to inflated promised pick-up times to customers and to a significant amount of idle capacity (i.e., "deadheading'", which is a truck moving without a load). In other words, in the absence of accurate information about the location of its own trucks, Schneider buffered itself using leadtime and capacity. By adopting its satellite-tracking system, Schneider was able to reduce both these buffers *and* offer improved delivery performance.

The **1st** Axiom of the IDIB Portfolio might be called the "trade-off axiom"; that is, by incurring increased cost for higher quality in one component of the IDIB Portfolio it should be possible to reduce the quality and cost of another component. *If the cost reduction is larger than the cost increase, then total portfolio cost has been reduced.* Further, as is often the case, some of the net savings can be invested in improving competitiveness (e.g., increasing fill-rates, reducing delivery leadtimes, offering higher levels of customization, reducing prices).

Other trade-offs are possible, too. For example, a make-to-stock manufacturer who has made the transition from a "push" management system (e.g., MRP) to a "pull" management system (e.g., JIT) has, everything else being equal, chosen to substitute buffer capacity for buffer inventory (and, often, to shift the inventory-buffering responsibilities to its suppliers and/or customers).

Sometimes, trade-offs are made to *increase* buffers. One example of this, experienced by every supply-chain manager who has replaced a domestic supplier with a lower-cost international supplier, is the increase in safety-stock inventory necessitated by the less-reliable transportation leadtimes from the off-shore supplier. In such cases, of course, the cost saving in implementation (i.e., delivery of items ordered from the sup-

[4] Assuming these processes are independent, the same safety stock will provide the same customer fill-rate, etc.

plier) isn't typically in transportation – indeed, transportation cost may increase – but in the cost of acquiring the materials themselves.

I believe that virtually all of the dramatic improvements reported by companies in managing their own internal supply chains or by companies and their partners in managing their shared supply chains can be interpreted – indeed, could have been forecast – using the **1st** Axiom of the IDIB Portfolio. That is, given the reduced cost of a "better" information system (as provided by innovations in information technology) and the already high cost of buffering, that information could be substituted for buffers – that is, information improved in quality and buffering reduced in quality[5] – *without* reducing customer service *and* at reduced total cost.

What the **1st** Axiom *doesn't* suggest is the magnitude of improvement in cash position, productivity, and competitiveness that so many partners have reported. This magnitude, of course, depends on the cost of the buffering required by being "blind" and how much of a buffer reduction (i.e., reduced inventory, capacity, and/or leadtime) a little "vision" provides.

So, will the future of supply-chain management involve even more extensive information-sharing or delegation of decision-making and implementation between supply-chain partners? The **2nd** Axiom suggests *some* more, but, depending on the partnership, perhaps not a great deal more.

The "**2nd** Axiom of the IDIB Portfolio" is this: ***Investment to improve the quality of any single component of an IDIB Portfolio will, over some range, decrease the total cost of the portfolio; but, beyond some quality level, increase the total cost of the portfolio.***

Although "axioms" are supposed to be self-evident, a little discussion of the **2nd** Axiom is appropriate. The **2nd** Axiom considers what happens to total portfolio cost if one varies the quality level of any single component of an IDIB Portfolio and adjusts the quality levels of the other three components to minimize the corresponding total portfolio cost. For example, consider varying the quality of implementation in some existing IDIB Portfolio. Applied to this example, the first half of the proposition is that if implementation is very low in quality, for example, that implementation leadtime has a large mean and variance, or that the accuracy of implementation is poor, then the total cost of

[5] Recall that, everything else being equal, reducing the amount of buffering reduces its capability to compensate for imperfections. Hence, reducing the amount of buffering reduces its "quality".

this IDIB Portfolio can be reduced by: (1) *increasing* the quality of implementation; and (2) *reducing* the quality of the corresponding information and/or decision-making systems, and/or reducing the amount of buffering (e.g., leadtime, capacity, or inventory).

The support for this proposition is that the cost of any of the four components of an IDIB Portfolio is an increasing and marginally-increasing function of its quality. In this instance, the proposition is that the cost of a low-quality implementation system is small, and that the cost to improve its quality is low compared to the high cost of the information, decision-making, and buffer systems *required to work in combination* with it. More specifically, given a low-quality implementation system, management would probably have been forced a very high-cost buffer system, possibly one with large inventories, or possibly one with large leadtime and/or capacity buffers. Hence, improving the quality of implementation – which should cost relatively little – and reducing the quality of its buffer system – which should save relatively more – will reduce total portfolio cost.

The second half of the proposition is that continuing to improve the quality of any single component and, correspondingly, decreasing the quality of one or more of the other three components of an IDIB Portfolio will, beyond some point, *increase* total portfolio cost. The support for this proposition, again, is that the cost of any of the four components of an IDIB Portfolio is an increasing and marginally-increasing function of its quality. In terms of the example, the proposition is that if the implementation system is already operating at a very high level of quality – e.g., nearly immediate, nearly perfect implementation – then the cost of any incremental improvement will be large relative to the savings generated by the corresponding decreases in the quality levels of buffering, information, or decision-making that this improvement in implementation facilitates.

The **2$^{\text{nd}}$** Axiom might be called the “golden mean” axiom; that is, it doesn’t make sense to invest too much in improving a single component of the IDIB Portfolio without making corresponding improvements in its other components.

7. The Future of Supply-Chain Management

Given that the current state-of-the-art in supply-chain management involves *some,* perhaps even a great deal, of information-sharing and *some* delegation of decision-making and implementation, the axioms suggest that even if the net savings from more information-sharing and delegation is positive, some *other* change in the IDIB Portfolio may yield *larger* net savings. In other words, the question is *not* whether more information-sharing or more delegated decision-making or more delegated implementation will reduce total portfolio cost. Instead, the question is: *Which changes in which components of the supply chain's IDIB Portfolio will facilitate the largest reduction in total portfolio cost?*

I believe that the most likely candidates for large cost savings are in *collaborative* decision-making and/or *collaborative* implementation. What's "collaborative"?

Broadly speaking, "collaborative" and "shared" mean the same thing, but I use the word "collaborative" in order to make an important distinction between visibility and participation. "*Shared* information" is about visibility: that is, within some given domain, all the partners "see" the same thing. Hence, "shared" decision-making or "shared" implementation might be interpreted to mean decision-making or implementation that is visible to all the partners. Although such visibility is important, *participating* in decision-making or *participating* in implementation means something much more significant. Participation means that both partners' objective functions, constraints, and relative capabilities are considered. From an operations-research perspective, collaborative decision-making and implementation involve *joint* optimization, not independent optimization.

The 2nd Axiom suggests that if partners are doing little or no collaborative decision-making or implementation, then it is possible that the greatest potential for improvement is in precisely these areas. Moreover, the higher the quality of the components of the supply chain's IDIB Portfolio is in *other* respects, the more likely shared decision-making and implementation are to provide the most significant total cost reductions.[6]

The IDIB Portfolio, *The Goal,* and the Theory of Constraints

In his ground-breaking book, *The Goal* (Goldratt and Cox, 1985), Eli Goldratt introduced the concept of bottlenecks in a firm's production capacity that limit its "throughput"; i.e., the rate at which a production system generates money through sales. From the perspective of the IDIB paradigm, throughput is a dollar-oriented quality characteristic of a firm's implementation (i.e., production) system. Correspondingly, Goldratt's bottleneck concept is that this quality characteristic – the capability of the implementation system to generate dollars through sales

[6] In addition, independent decision-making based on the same information, or delegated decision-making and implementation, *at best,* yield locally-optimal decisions and actions.

– is limited by the bottleneck process(es) internal to the implementation system.

Of course, a firm's implementation system doesn't generate throughput all by itself. Decisions must be made about what to produce, when to produce it, etc. This is the role of the decision-making system. And, in order to make decisions, the decision-making system requires information that's provided by the information system. Finally, buffer systems are there to provide throughput by compensating for whatever imperfections might exist in the IDI_ systems. In other words, it is a firm's *entire* IDIB Portfolio that generates money through sales, not just its implementation system.

Although Goldratt doesn't recognize the IDIB Portfolio explicitly, his prescriptions most certainly apply to it. For example, Goldratt prescribes that decision-making in the management system should be focused on the "drum" of the implementation-system's bottleneck(s) and the "ropes" that feed it. Goldratt also has prescriptions about the form, location, and amount of buffering that should be provided[7], and about the nature of the information systems[8] management should use in decision-making. In brief, Goldratt recommends that all four components of a firm's IDIB Portfolio should be focused on the supply (and demand) bottlenecks. Hence, the IDIB paradigm and the bottleneck paradigm are consistent, indeed, complementary, to one another.

Specifically, I believe that the IDIB paradigm enriches the paradigm of *The Goal* in several ways. For example, the IDIB paradigm suggests that the "goal" of making money can be achieved by virtually unlimited number of different IDIB Portfolios, each component contributing in its characteristic way (i.e., gathering information, making decisions, etc.), each imperfect in different ways, and each compensating for imperfections in the others. Second, of course, that the quality of each of these components is a decision variable.

Next, recall that Goldratt points out that the only guaranteed way for a firm to make money is to simultaneously increase throughput and reduce "inventory" and "operating expenses".[9] However, Goldratt offers relatively little guidance about *how* to reduce inventory and operating expense, much less minimize them. The IDIB paradigm suggests how: select the least total-cost IDIB Portfolio. In other words, since most of a

[7] For example, buffer inventory in front of bottlenecks and buffer leadtimes to protect delivery dates.

[8] See Goldratt, 1991.

[9] According to Goldratt, "inventory" is the dollars that a firm has invested in things that it intends to sell, while operating expense is the cost of things that the firm does to turn inventory into throughput.

firm's operating expenses are driven by management's chosen IDI_ systems and most of its inventory is in its safety-stock and capacity buffers, by carefully choosing its IDIB Portfolio, management will minimize its combined operating expenses and inventory

Finally, note that the "bottleneck" concept is *imbedded* in the **2nd** Axiom. Here's how: Consider a company with medium-to-high quality decision-making and implementation systems, but a low-quality information system. Under these circumstances, the overall quality of this company's IDI_ systems is limited by its low-quality information-system. According to the **2nd** Axiom, if this company invests in a better information system, the overall quality of its IDI_ systems will improve. This improvement facilitates the reduction in its associated buffers. Given some (low) range of information-system quality, the effect should be to reduce the total cost of this company's IDIB Portfolio.

Next, consider a company with an excellent implementation system, say a state-of-the-art cellular system, but with mediocre-quality information and decision-making systems. Additional improvement in this company's implementation system will yield an improvement in its overall IDI_ systems, and facilitate a reduction of the corresponding buffers. However, the money saved on the buffer system may be less than the additional cost of the improved implementation system. The result is an increase in total portfolio cost.

From Goldratt's viewpoint, the "bottleneck" in the quality of first company's IDI_ systems is its information system: a dollar invested there yielded more than a dollar saved on the buffer system, thereby reducing the firm's total inventory and operating expenses. On the other hand, the "bottleneck" in the quality of second company's IDI_ systems *wasn't* its implementation system. Hence, a dollar invested there is a dollar wasted.[10]

The Future is Wow for Some Supply-Chain Partners

Some supply-chain partners are already sharing decision-making. Since 1995, Heineken USA, Inc. and its independent distributors have been sharing information and decision-making about the replenishment of Heineken's beer products under a system called HOPS: the Heineken Operational Planning System (http://64.158.250.111/news/heineken.

[10] Of course, neither of these improvements, whether they would reduce its total IDIB Portfolio cost or not, necessarily increases the company's throughput. Throughput would only be increased if the changes to the entire IDIB Portfolio increase availability or otherwise make the company's products more competitive. Nonetheless, even if throughput isn't increased, reducing the total IDIB Portfolio cost reduces a company's operating expense, which increases profits; i.e., makes more money.

html and http://64.158.250.11l/news/archive99/06091999.html). Intel and its customer computer-assemblers (e.g., IBM, Dell, Compaq) have been using a collaborative information- sharing and decision-making system to manage the assembler's inventories of computer chips under Intel's Supply-Line Management (SLM) program.

One well–known and widely-implemented system for information-sharing and collaborative decision-making in supply chains is Collaborative Planning, Forecasting, and Replenishment (CPFR).

8. Collaborative Planning, Forecasting, and Replenishment (CPFR)

CPFR is a process model, shared by a buyer and supplier, through which inventory-status, forecast-, and promotion-oriented information are shared and replenishment decisions are made. In IDIB Portfolio terms, a process model for sharing information and decision-making between a buyer and supplier.

CPFR began with a pilot program between Wal-Mart and Warner-Lambert, called CFAR: "Collaborative Forecasting and Replenishment". In 1997, the Voluntary Interindustry Commerce Standards (VICS) Association (http://vics.org) developed the "CPFR Initiative" (http://www.cpfr.org). In 1998, VICS published the first "CPFR Guidelines" (http://www.cpfr.org/Guidelines.html). Since then, a large number of partners have developed CPFR pilots. The appendix provides a partial list. Several partnerships have subsequently adopted CPFR as a standard way of doing business with one another.

The CPFR Process

CPFR consists of 9 process steps, as follows:

Step 1. Develop Front-End Agreement: Roles, Measurement, Readiness

Step 2. Create Joint Business Plan: Strategies and Tactics

Step 3. Create Sales Forecasts: Buyer and supplier both create customer-demand forecasts

Step 4. Identify Exceptions in Sales Forecasts

Step 5. Resolve Exceptions: Agree on single forecast or agree to disagree

Step 6. Create Order Forecasts: Buyer and supplier both create plans for buyer orders

Step 7. Identify Exceptions in Order Forecasts

Step 8. Resolve Exceptions: Agree on single plan for buyer orders

Step 9. Order Generation

More details about these steps and the roles of the buyer and supplier in each step are provided in Figure 11.1 and at the VICS CPFR website (`http://www.cpfr.org`). However, the basics of CPFR are straightforward: First, the partners share information about demand. If the buyer is a retailer – and so far, most buyers using CPFR *are* retailers, then demand is retail customer demand. If the buyer is a manufacturer or assembler then demand is generated by the manufacturer or assembler's trial master-production schedule. Then, significant differences between the buyer's and seller's demand forecast, labeled "exceptions", are discussed and resolved. These are Steps 3-5 above. Then, buyer and supplier share plans for orders that the buyer will place with the supplier, based on the shared demand forecasts. Again, exceptions are identified and resolved (Steps 6-8). Subsequently, using the shared order plan, actual orders are generated (Step 9). The foundation for Steps 3-9 is the so-called "front-end agreement", under which the roles of the buyer and supplier, and their capabilities to perform these roles are assessed. In this step, targeted performance and measures are also adopted. In Step 2, strategies and tactics are specified in detail.

The benefits reported by CPFR partners, as might be predicted by the axioms, are increased inventory turns (i.e., lower buffer inventory) and increased fill-rates for the SKU's involved; that is, higher levels of customer service.

Several consulting firms offer software systems and support for CPFR, among them Logility, Inc. (`http://www.logility.com`) and Syncra Systems, Inc. (`http://www.syncrasystems.com`). CPFR is also being implemented on B2B exchanges such as Worldwide Retail Net, Transora, and NetXchange.

Based on the success of CPFR between single buyer-supplier pairs, thought leaders in CPFR have suggested its extension to include collaboration with the carrier that transports goods between the buyer and supplier. This is called "CTM": Collaborative Transportation Management (`http://www.cpfr.org/WhitePapers/CTMwhitepaper.pdf`). It has also been suggested that in order to be truly successful, collaboration should involve all of the links of a supply chain, under a scheme labeled "n-Tier Collaboration" (`http://www.cpfr.org/WhitePapers/nTierProposal.doc`). Under such a scheme, for example, not only would Kimberly-Clark collaborate with Wal-Mart to plan Wal-Mart's orders of Kimberly-Clark products, but Kimberly-Clark would, in turn,

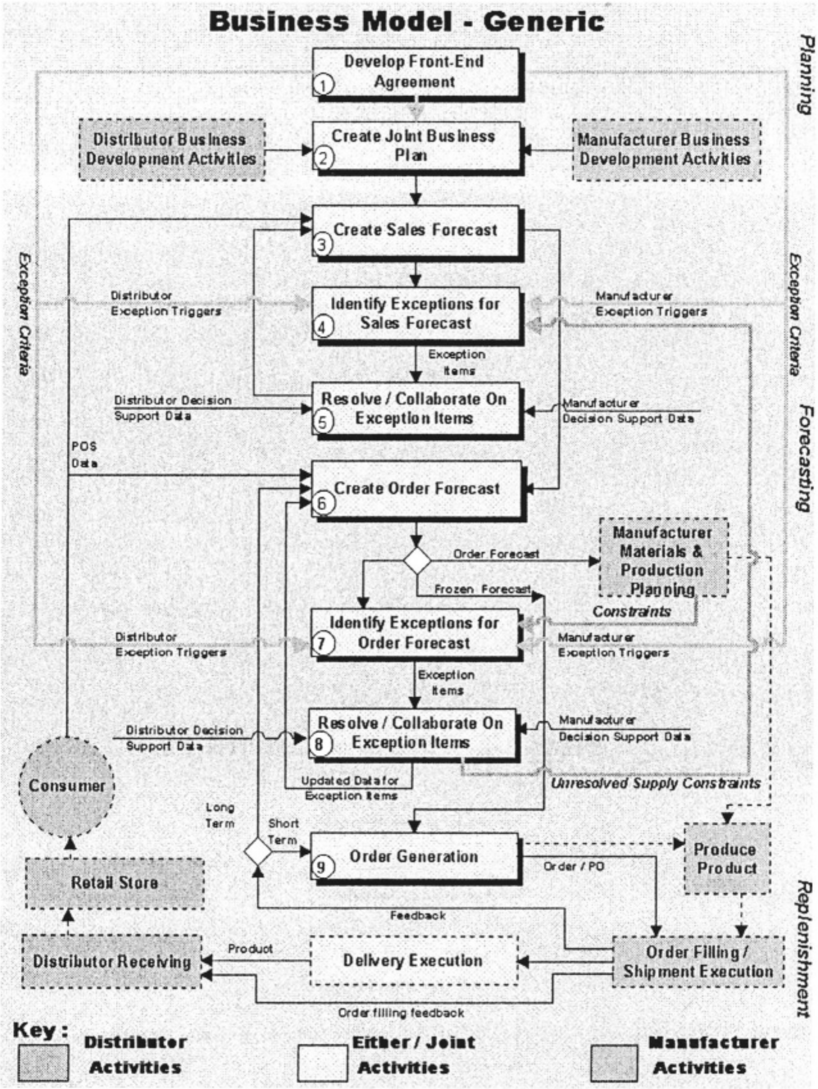

Business Model - Generic
Planning
Forecasting
Replenishment
Exception Criteria
Exception Criteria
1 Develop Front-End Agreement
Distributor Business Development Activities
2 Create Joint Business Plan
Manufacturer Business Development Activities
3 Create Sales Forecast
Distributor Exception Triggers
4 Identify Exceptions for Sales Forecast
Manufacturer Exception Triggers
Exception Items
Distributor Decision Support Data
5 Resolve / Collaborate On Exception Items
Manufacturer Decision Support Data
POS Data
6 Create Order Forecast
Order Forecast
Manufacturer Materials & Production Planning
Frozen Forecast
Constraints
Distributor Exception Triggers
7 Identify Exceptions for Order Forecast
Manufacturer Exception Triggers
Exception Items
Distributor Decision Support Data
8 Resolve / Collaborate On Exception Items
Manufacturer Decision Support Data
Consumer
Updated Data for Exception Items
Unresolved Supply Constraints
Long Term
Short Term
9 Order Generation
Order / PO
Produce Product
Retail Store
Feedback
Product
Distributor Receiving
Delivery Execution
Order Filling / Shipment Execution
Order filling feedback
Key:
Distributor Activities
Either / Joint Activities
Manufacturer Activities

Figure 11.1. The 9-Step CPFR Process.

collaborate with its suppliers to determine Kimberly-Clark's orders for materials for its paper products. So, in effect, Wal-Mart and Kimberly-Clark's suppliers would be collaborating with Kimberly-Clark to determine plans to supply and order Kimberly-Clark's paper products.

The Challenges of CPFR

Although CPFR has enormous potential for reducing the total cost of any supply chain's IDIB Portfolio, there are also enormous challenges. At the most fundamental level, buyers and suppliers must develop trust that each will treat the other fairly and honestly. Prerequisite for this are incentives to do so. Again, see Cachon, 2004 for a review of the supply-chain literature on the management of incentive conflicts with contracts. On a technical level, buyers and suppliers must develop a common language for identifying products and making decisions about them. (See `http://www.cpfr.org/WhitePapers/CollaborationDataModelingA.pdf`). Similarly, systems must be developed for linking the buyer's and supplier's business processes. This will involve a great deal of system change and training. Third, security protocols must be implemented that will safeguard both partners from leaks of proprietary information.

Nonetheless, I believe that buyers and suppliers who find ways to overcome these challenges will achieve a competitive advantage, particularly for products that complete primarily on price and availability (e.g., consumer products). This advantage will either force the competitors of CPFR partners to adopt similar techniques or force them out of business.

9. Research Topics in CPFR

CPFR poses many interesting questions for supply-chain researchers, questions whose answers involve models that have yet to be developed.

At the broadest level, there are questions involving the "drivers" for collaboration. For example, what are characteristics of buyers, suppliers, and the environments in which they operate that promotes a desire on either of their parts to collaborate using CPFR? It is well known that agency losses occur in decentralized supply chains that involve hidden information and/or hidden actions. What is less well known are the circumstances under which *both* the buyer and supplier will be better off by collaborating. See, for example, Monahan, 1984; Weng, 1995 and Taylor, 2001. In the absence of these circumstances or other incentives for it, collaboration, in general, and CPFR, in particular, seem doomed.

With respect to CPFR in particular, consider a buyer-supplier pair. Among the most interesting research questions are: First, how should the "front-end" agreement be structured in order to maximize – or merely, increase – the profits for both partners? What role will each partner play? How should performance improvement be measured? Perhaps most important: how should the benefits of improved performance be shared between the buyer and the supplier so that both will be better off?

Another set of interesting questions involves defining the elements of the data to be shared. For example, given the cost of data processing and security considerations, should SKU- level data be shared or should only aggregate information be shared? What aggregation/disaggregation procedures are best?

Finally, given that there are costs and benefits associated with exception-processing, regardless of *how* they are processed, how should exceptions be defined? More fundamentally, what does it mean to process an exception; that is, given a significant difference between and the buyer's and seller's forecast (planned orders), how should the difference be resolved?

10. Research Topics Related to Supply-Chain Collaboration

Another set of interesting research questions involves the examination of supply-chain collaboration in general, whether or not the tools used are those of CPFR. For example, given a supply chain with some given level of collaboration in information-sharing, decision-making, implementation, and buffering – including no collaboration whatsoever – then, *assuming* that collaboration will be increased, *which links* should be involved, and *in which components* of the IDIB Portfolio will collaboration yield the largest payoffs to the partners involved and to the entire supply chain? Further, how will be benefits of collaboration be measured and shared among partners and along the chain?

An equally interesting set of questions involves collaboration itself. For example, consider a supply chain with some given level of collaboration in information-sharing, decision-making, implementation, and buffering – including no collaboration whatsoever. Then, given the costs and challenges of collaboration, is collaboration the most cost-effective way to improve supply-chain performance? In particular, improving the quality of one or more of the components of one of the partner's IDIB Portfolios may yield larger improvements at lower cost. If so, then, again, how will benefits be measured and shared?

11. Research on the IDIB Portfolio

In generalizing the OR paradigm, the IDIB Portfolio paradigm provides rich and challenging opportunities for operations researchers.

I believe that the OR paradigm has focused almost entirely on modeling only a single component of supply-chain IDIB Portfolios, typically on the decision-making component. To the extent that the associated information and implementation systems are represented at all, these models treat them as fixed, as costs, or as constraints, but *not* as decision-variables. For example, such models typically take the quality of the information system as given (e.g., demand is known or given by some particular probability distribution with fixed parameters) and implementation systems are represented as parameterized leadtimes or by their cost drivers (e.g., the parameterized cost of performing a set up or the unit cost of holding inventory). The *form* of buffering (i.e., inventory, leadtime, or capacity) is usually also fixed. In general, the objective function is to select the decision (or decision-rule) that minimizes the associated implementation costs plus the too-much and not-enough costs associated with the buffer system. Other models, in particular, those that treat either the information system or the implementation system as a decision variable, typically ignore the *other* three components of the IDIB Portfolio as decision variables.

That's the "bad news".

The "good news" is that the OR paradigm's focus has facilitated the development of fairly sophisticated decision rules and provided limited insight into information and implementation systems. It is also "good news" that many of the techniques employed in the OR paradigm, for example, well-proven estimation and optimization techniques, can be applied to improving, if not optimizing, the IDIB Portfolio model of a management system. For example, cost-estimating techniques that are already being applied to estimating the cost of a given implementation system can be applied to determining the cost of *alternative* information, implementation, and buffer systems.

Notwithstanding the availability of OR tools, the challenges of IDIB Portfolio optimization, even if done heuristically, are daunting. Recall, the "optimal IDIB Portfolio" is the portfolio of information, decision-making, implementation, and buffering whose combined cost *plus* the cost of failing to provide precisely was demanded or required (e.g., back-ordering or expediting cost) is the minimum possible cost. Hence, IDIB Portfolio optimization involves *four* decision-variables, not simply one. Further, in order to be useful in selecting an IDIB Portfolio, IDIB models *must* be able to capture complex cost-quality interactions *among* these components, not just the cost-quality characteristic of each component.

Practice-oriented research also presents interesting research opportunities. For example, consider a firm's current IDIB Portfolio and assume management has the desire to "improve" it. The question is: Where (i.e., which component) is the "bottleneck" of the current IDI_ systems and what improvement should be made to it? *Before* this question can be answered, operations researcher must first develop techniques and measures to assess the capabilities of each of its components. Next, in order to make the most cost-effective improvement, managers will require *a priori* estimates of the marginal costs and benefits of the *next* increment in quality in their current components. Further, there are questions relating to the timing of IDIB Portfolio changes. For example, although adopting current technology might reduce total portfolio cost, should management forego adopting it, and, instead, wait for new technology that might provide even larger total cost reductions? Technology forecasting must play a role in answering this question.

Despite the daunting nature of the challenges to improving a firm's current IDIB Portfolio or optimizing a proposed IDIB portfolio, operations researchers will be foregoing a tremendous opportunity if they continue to focus on optimizing one component of the IDIB portfolio at a time. Worse, they commit the Cardinal Sin of Operations Research: *Sub*-Optimization.

12. Some Related Literature

Most of the literature in supply-chain management can be viewed from the perspective of the IDIB Portfolio. As explained above, I believe that most of this literature examines the choice of the decision (e.g., order quantity, target inventory) or decision-rule (e.g., EOQ) to minimize total too-much and not-enough costs. See the discussion of the newsvendor model above, for example. Some of this literature also examines the impact of implementation issues on decision-making. For example, in a highly-stylized way the EOQ formula represents the impact of set-up (i.e., implementation) cost on the optimal order quantity. However, there are some signs of interesting developments.

A growing body of literature examines the value of information-sharing in managing a supply chain. Lee et al., 2000, for example, have shown that information-sharing can dampen the so-called "bullwhip effect" so often observed in supply chains. Chen et al., 2000, have shown that the bullwhip effect can be reduced by centralizing information. Cachon and Fisher, 2000, study the value of sharing demand and inventory-level information in a supply chain. Aviv, 2001, examines the effect of collaborative forecasting on supply-chain performance.

A small, but growing body of research involves collaborative replenishment decision-making to take advantage of shared information. For example, Song and Zipkin, 1996, develop an inventory-replenishment policy to take advantage of information about supply conditions. Aviv, 2002, examines joint forecasting and replenishment processes. Iyer and Ye, 2000, develop a model to assess the value of information-sharing in a retail environment in which retailers share promotional information with their suppliers.

However, I am not aware of *any* research that addresses itself to "managing the IDIB Portfolio"; that is, making decisions about the nature and quality of *all* its four components: the information, decision-making, implementation, and buffer systems.

Nonetheless, there are two papers of particular relevance to "managing the IDIB Portfolio" that I would like to draw attention to:

Milgrom and Roberts, 1988, develop a model of a manufacturer and the market for its products, and, for this particular model, establish that the optimal management system will either produce to inventory or produce to customer order. In IDIB Portfolio terms, Milgrom and Roberts examine a model in which there is a cost to acquire information about market demand and a cost to produce. Their analysis concludes that a profit-maximizing firm will either: (1) acquire no additional information about customer demand and produce entirely to inventory; or (2) acquire complete information about customer demand and produce entirely to customer order.

It should be noted that Milgrom and Robert's result is specific to the assumptions of the model they propose. Other assumptions would yield different results. For example, Milgrom and Robert's model ignores the production (implementation) leadtime, customer preferences, and competition. Given significant production leadtimes and customer preference for availability over variety, a manufacturer that might otherwise choose make-to-order might be forced to make to stock. Alternatively, given customer preference for variety over availability, a manufacturer that might otherwise choose make-to-stock might be forced to make to order. Nonetheless, Milgrom and Roberts are the first to suggest tradeoffs between information and inventory (i.e., one form of buffer).

Hariharan and Zipkin, 1995 implicitly suggest tradeoffs between the characteristics of a company's information system and implementation system. In particular, within a supply-chain setting, they establish the equivalence between the information system's ability to "see" one more (less) period of demand into the future and the ability of the implementation system to produce one period faster (slower).

The use of the "ICB Portfolio" paradigm, which is closely related to the IDIB paradigm, in teaching operations management can be found in Schwarz, 1998.

13. Summary

In this chapter, I have introduced a new paradigm for management in general and for managing supply chains, in particular, called the IDIB Portfolio. The IDIB Portfolio, which can be viewed as a generalization of the OR paradigm, takes the view that the quality of *all* four components of a management system – not just the buffer system quality (i.e., size) – are decision variables; and that the underlying quality-cost function for each of these components – that is, the cost to move them in the direction of perfection – is increasing and marginally increasing. Finally, that there is a cost associated with the entire IDIB Portfolio – again, not just the buffer system – for failing to provide whatever is required or demanded. Hence, the "optimal" IDIB Portfolio is the portfolio that minimizes the total costs of all of its components *plus* the cost of failing to provide whatever is required or demanded.

The IDIB Portfolio and its axioms provide insight into the evolution of supply-chain practice to date, and, I believe, suggest that the future of supply-chain management practice will involve a significant level of collaborative decision-making and implementation: a level of collaboration in decision-making and implementation that is comparable to the level of information-sharing in contemporary supply-chain management. I have described the VICS CPFR initiative as one example of this future. Finally, I have suggested several research topics involving CPFR and, more broadly, challenging new research into the IDIB Portfolio paradigm.

Acknowledgement

I am indebted to two anonymous reviewers for their helpful comments on an earlier draft. I am particularly indebted to Nick Petruzzi for many insightful suggestions, including the notion that the IDIB paradigm generalizes the OR paradigm and the links between the IDIB Portfolio paradigm and Goldratt's *Goal*-oriented paradigm.

Appendix: List of Buyers and Suppliers participating in CPFR partnerships

Source: The VICS CPFR® Matrix, January 2002. This list is maintained monthly by MoonWatch Media, Inc. See http://www.retailsystems.com/communitycenters/cccc/cpfrmatrix.pdf.

Table 11.A.1. *Buyers*

10 Internal Affiliates
4 Retailers
850 n-Tier Partners
Ace Hardware
Albertson's
Best Buy
Canadian Tire
CVS
Dansk
Dealers
Delhaize le Lion
Distributors
Do It Best
Eckerd
Federated Department Stores
H.E. Butt
Home Depot
J.C. Penney
Jusco
Londis
Marshall Field's
Match Supermarket
McDonald's
McDonald's France
Meijer
Mervyn's
Radio Shack
RiteAid
Royal Ahold
RONA
Safeway
Safeway (UK)
Sainsbury
SAKS
Sears Roebuck
Somerfield
Sports Authority
Staples
Superdrug
Target
Tesco
TruValue
Walgreens
Wal-Mart
Wickes Furniture
Woolworth UK

Table 11.A.2. Suppliers

12 Suppliers
20+ Suppliers
Ashley Furniture
Ball Sports
Black & Decker
Broyhill
Chanel
Chapin
Colgate-Palmolive
Compaq
Eastman Chemicals
ECPG3
Eli Lily
Feather Fruit Growers' Cooperative
FujiFilm
GE Appliances
General Mills
Genovs
Georgia Pacific
Harley-Davidson
Hasbro
Heineken
Henkel
Herlitz
Hewlett-Packard
HYKo
Inland Paperboard & Packaging
International Paper
John Deere
Johnson & Johnson
Kao
Kimberly Clark
Kraft
Lever-fabrege
Levi
Levi Strauss
Liquid Nails
Liz Claiborne
Manco
Mars
Master Lock
Meriat
Mitsubishi Motors
Nelson
Nestle UK
New Balance
Pacific Coast
Panasonic
Philips Consumer
Pillowtex
PlumbPak
Polo Ralph Lauren
Proctor & Gamble
Reynolds Metals
Rowe Companies
Sara Lee
Schering-Plough
Solo Cup
Spectrum
Thomson Electronics
Timberland
Truya
Unilever Argentina
Vandemoortele of Belgium
Warner-Lambert
Whitehall Robbins
Woodstream
YKK

References

Aviv, Y. (2001). The effect of collaborative forecasting on supply-chain performance. *Management Science,* 47(10): 1326–1343.

Aviv, Y. (2002). Gaining benefits from joint forecasting and replenishment processes: The case of auto-correlated demand. *Manufacturing & Service Operations Management,* 4(1): 1–18.

Cachon, G.P. (2004). Supply chain coordination with contracts. In Kok, A.G. de, and Graves, S.C., editors, *Handbooks in Operations Research* *and Management Science: Supply Chain Management.* North-Holland, Amsterdam, The Netherlands. Forthcoming.

Cachon, G.P. and Fisher, M. (2000). Supply-chain inventory management and the value of shared information. *Management Science,* 46(8): 1032–1050.

Çetinkaya, S. and Lee, C. (2000). Stock replenishment and shipment scheduling for vendor-managed inventory systems. *Management Science,* 46:217–232.

Chen, F., Drezner, Z., Ryan, J.K., and Simchi-Levi, D. (2000). Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, leadtimes, and information. *Management Science,* 46(3):436–443.

Ehrhardt, R. and Taube, L. (1987). An inventory model with random replenishment quantities. *International Journal of Production Research,* 25:1795–1804.

Fiorito, S.S., May, E.G., and Straughn, K. (1994). Quick response in retailing: Components and implementation. *International Journal of Retail and Distribution Management,* 23:12–21.

Goldratt, E. (1991). *The Haystack Syndrome: Sifting Information Out of the Data Ocean.* North-River Press, Groton-on-Hudson, New York.

Goldratt, E. and Cox, J. (1985). *The Goal.* North-River Press, Groton-on-Hudson, New York.

Hariharan, R. and Zipkin, P. (1995). Customer-order information, leadtimes, and inventories. *Management Science,* 41:1599–1607.

Iyer, A.V. and Ye, J. (2000). Assessing the value of information-sharing in a promotional retail environment. *Manufacturing & Service Operations Management,* 2(1):128–143.

Karlin, S. (1958). One-stage models with uncertainty. In Arrow, K.J., Karlin, S., and Scarf, H., editors, *Studies in the Mathematical Theory of Inventory and Production,* chapter 8. Stanford University Press, Stanford, California.

Lee, H., Padmanabhan, V., and Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science,* 43:546–558.

Lee, H., So, K.C., and Tang, C.S. (2000). The value of information-sharing in a two-level supply chain. *Management Science,* 46(5):626–643.

Lowe, T. and Schwarz, L.B. (1983). Parameter estimation for the EOQ lot-size model: Minimax and expected value choices. *Naval Research Logistics,* 30(2):367–376.

Milgrom, P. and Roberts, J. (1988). Communication and inventory as substitutes in organizing production. *Scandinavian Journal of Economics,* pages 275–289.

Monahan, J.P. (1984). A quantity-discount pricing model to increase vendor profits. *Management Science,* 30:720–726.

Schwarz, L.B. (1998). A new teaching paradigm: The information/control /buffer portfolio. *Production and Operations Management,* 7:125–131.

Song, J.S. and Zipkin, P.H. (1996). Inventory control with information about supply conditions. *Management Science,* 42(10):1409–1419.

Taylor, T.A. (2001). Channel coordination under price promotion, midlife returns, and end-of-life returns in dynamic markets. *Management Science,* 47:1220–1234.

Weng, Z.K. (1995). Channel coordination and quantity discounts. *Management Science,* 41:1509–1522.

Yano, C. and Lee, H. (1995). Lot sizing with random yields: A review. *Operations Research,* 43:311–334.

This page intentionally left blank

Chapter 12

MYTHS AND REALITY OF SUPPLY CHAIN MANAGEMENT: IMPLICATIONS FOR INDUSTRY-UNIVERSITY RELATIONSHIPS

André Kuper
Hewlett-Packard Company
Palo Alto, California
andre.kuper@hp.com

Sarbani Bublu Thakur-Weigold
Hewlett-Packard Company
Ratingen, Germany
sarbani.thakur-weigold@hp.com

Abstract Current surveys reveal that the practice of Supply Chain Management has had little impact upon business performance. This opportunity loss stands in stark contrast to the wealth of knowledge, and volumes of materials available to both researchers and practitioners alike. The dissemination of Supply Chain Management theory, from its roots in Operations Research to its application in business operations, is clearly dysfunctional. Root causes of faulty exchange between universities and industry include unspoken assumptions, which, one the one hand facilitate research, but may sabotage solution implementation if not properly contextualized onsite. Universities and industry speak different languages, address different audiences, have their own mutually-incompatible objectives, and divergent turn-around rates for innovation. Nevertheless, they remain vitally interdependent in the fields of Supply Chain Management and Operations Research. This chapter discusses how the misunderstandings between these parties perpetuate a number of dangerous myths of operations management, all of which are exacerbated by media reductions. An unforgiving business reality demands that industry and academia establish an ongoing dialogue to rise to increasingly complex supply chain management challenges. This dialogue has to evolve from sporadic individual partnerships, to a genuine, systematic collaboration. The examples illustrate the need to validate and test research models in real operational environments, to ultimately develop solutions which can be implemented without peril. Furthermore, the cases are the outcome of management training programs which have traditionally devalued operations management when defining the core competencies needed by decision-makers and researchers alike. This lack of priority goes hand-in-hand with the over-simplification and distortion of research, which dangerously percolates down to its ostensible users. Reflecting upon the experience in Supply Chain Management practice at HP, including its successful history of academic partnerships, the authors identify a number of success factors for fruitful collaboration.

1. Introduction

The current business environment compels managers and researchers alike to focus on asset management. The manufacturing industry's ongoing reliance on supply chain partners to realize economic value has created a global network of dependencies, in which the overall performance is dictated by the weakest link. Moreover, the inexorable evolution of these networks has shifted attention from fixed assets (like manufacturing equipment and buildings), to variable assets (like inventory). It will come as no surprise that supply chain management – the science and practice of end-to-end flows – has emerged as a determining factor in business success.

Successful adaptation begins when these shifts are recognized by all stakeholders. Here we address two particular constituencies, those who reflect upon the complex phenomena of supply chains, and those who must manage, and bear the brunt of their quirks. On the one hand, today's business schools and universities face the challenge of appropriately directing investment into the most promising research projects. Their intellectual output will eventually reach a wider audience, however, often with unintended consequences. Seized upon by an insatiable industry of consulting methodologies and management literature, this intellectual property often arrives at the "user" in unrecognizable form. The solutions observed in one particular success story are rapidly packaged into repeatable initiatives or programs, regardless of the business case at hand.[1] On the other hand, managers in industry face mounting pressure to validate the theories and slogans upon which they base fateful operational decisions. The volume of both scholarly and publications and their (often distorted) popular versions is daunting: how to translate all these ideas into operational success?

The reality of supply chain management contradicts one of the oldest beliefs shared by both of these parties: with proper management, companies can control their own growth. This assumption seems self-evident and is so fundamental to the *raison d'être* of MBA training, that it no longer merits discussion. However, a recent interview with the chip company LSI Logic shows that even this apparently unshakeable assumption requires revision: their management states that they cannot afford to build a new production facility on their own.[2] Imagine training a future executive to deal with dependency, volatility, and lack of knowledge. Contemporary cases like this one abound, mounting evidence that both the practice and theory of operations management could benefit from critical self-examination.

What follows is a symptomatic listing of the most common assumptions that we, at HP, have observed in academic (and subsequently), popular thinking. For those who must face reality daily, the persistence of misrepresentations and myths in supply chain management is striking, not least because there can be no doubt that each of the topics we sketch here is the subject of sustained and serious reflection. This leads us to suggest that something is dysfunctional in the dissemination of academic research into the business world – insights upon which real managers can successfully act.[3]

The rupture in this feedback loop is an absurd irony. When asked, scholars will unabashedly confess that the validation of their knowledge in practice is not even an objective; profound and rigorous thought leadership is the "real" goal. When reading an article on management science, one must ask which audience the author implicitly addresses. Can universities be moved to seek more than the self-referential approval of their erudite peers? Is the quality of an innovative MBA course in operations management to be judged by other professors – or might operations managers have input on how to reach the ideal operational state? Can the sometimes-tedious labor of applying knowledge to organizational situations become the subject of intellectual passion? We firmly believe that the answer is yes. The consequences of the contemporary, mutually dismissive estrangement are costly and avoidable: mistaking the theoretical model for reality remains a too-common pitfall in a world where expertise is available in overwhelming abundance.

Our objective is to share what we at Hewlett Packard's "Innovation Diffusion" team have learned in dozens of business-transforming Supply Chain Management projects. This chapter will discuss how to validate theory with business reality and how, in turn, to deploy academic research[4] to better interpret – and thereby survive – the complexities of business change. It will conclude with a vision of profitable collaboration between the doers and thinkers in management science, drawn from the positive experiences HP has had over the years.

Today's supply chain challenges

Cost is the incessant concern of many managers today, especially in high-tech manufacturing. Following euphoric decades of double-digit growth, we find ourselves in a period of reset revenue expectations. In this industry, the combined factors of stalled growth, high debt load, and the pressure to transform existing business models, demand that cost be eliminated wherever possible. There is an inherent paradox in the challenges that arise, since cost-cutting alone will not necessarily have any meaningful impact, and, simultaneously, the trough is usually the right time for strategic investment. For management, shrewd and swift investment can become agonizingly complex, since the assumptions upon which existing business models were built, are increasingly out of date. Suddenly, business-as-usual may not be an option.[5] The paradox that our high-tech industry must wrestle with is that if we just cut costs without changing the business fundamentals, we may ignore the realities that made cost an issue in the first place. In the medium- to long-term, cost-cutting, which tends to short-sighted and motivated by immediate returns, must be associated with risk. What we jettison, outsource, sell, and lay-off may well inhibit revenue and profit growth when the market picks up.

For supply chain managers, we face specific risks, with potentially devastating consequences. Ignoring the fact that increased volatility in demand will actually require more inventory – not less[6] – means that cutting stock in the hopes of freeing up capital will ultimately impede sales. In commodity markets like for consumer computers, availability is as critical as low price: a lost sale is a lost customer and bundled with peripheral products, can mean missing years of follow-up revenue. Considering a manufacturer's inherent dependency upon up- and downstream partners, the self-centered habit of passing on costs (in the form of inventory), also weakens our end-to-end supply chains. Typically, these short-term actions will increase the total costs for every party involved, while reducing flexibility. Both outcomes will negatively impact revenue and profitability. Even worse, it might drive partners out of business, and ultimately, destroy a market.

Such can be the consequences of the well-meaning, but poorly-informed, intuitive response. It should demonstrate the urgency of research and business investment into one of our high-tech industry's most pressing topics, the phenomenon of commoditization. Two trends contribute to the commoditization of goods: customer education, and specialization. When customers do not honor the value of a higher quality or better-engineered product by paying more, they render it a commodity. As a consequence, competition for revenue will be based on supply chain attributes like availability and price. Similarly, increased industry specialization has driven the use of commodity platforms like storage, data processing, input devices, displays, and software. This eliminated the possibility of leapfrogging the competition on innovative features, and left players competing on time-to-revenue by pushing generic products with few distinguishing features. In this scenario, in order to compete in commodity markets the supply chains need to be agile, responsive, and create velocity. At the same time, traditional competitive aspects like capacity utilization and economies of scale will, in turn, diminish in importance. In this context, capacity utilization and economies of scale actually reflect a dangerously obsolete assumption: that effective management of fixed assets is a driver for business success. If not critically reappraised, our favorite levers may become constraints to profitability in markets that grant players no second chance.[7]

Let us proceed to another subtlety that must be managed in the commodity space. A vendor's ability to meet demand is determined by its cash constraints and by dependencies in the supply chain. In commodity environments, the ramp rates and flexibility of volume are critical to a successful response to demand volatility, and thus the prerequisites to profitability at HP. Our research[8] has shown that product ramp rates are constrained by how effectively upstream suppliers obtain access to cash. These financial positions are fairly static, requiring vendors to nurture the health of their own suppliers.[9] An example, the strategic silicon market, whose players are critical suppliers to high-tech manufacturers, demonstrates that industry consolidation is another threat to assurance of supply. In this extremely capital-intensive industry, once, in the zeal of cost reduction, capacity has been cut back, it takes a year or more to install new fabs.[10] Similar trends are discernible in packaging, logistics, plastics, and contract manufacturing.

The common element in all the complex phenomena discussed here is relentless pace of change, and more anti-intuitive dependencies than one might expect. If the boundary conditions of our operations are constantly being revised, any reliance upon benchmarks and successful behavior from the past becomes infeasible. Since the intuitively sound approaches that worked in the past suddenly fail, even experienced managers find themselves at risk. The lack of reliable learning that can be applied to current conditions is also a wake-up call for operations research: an attention to cross-functional, dynamic environments has become critical to economic growth.

Inter-disciplinary research spanning more than two fields is, however, untypical to university projects (whose internal metrics reward specialization and eschew the apparent dilution of disciplines).[11] To illustrate this let's turn to the topic of capacity utilization. A recent study from Arizona State University[12] claimed that maximum capacity utilization is the right approach for profitable automotive manufacturing. Not only did this theoretical position not raise an academic eyebrow, it is apparently reflected in current U.S. automotive manufacturing. Carmakers use steep discounts to sell off the mountains of unsold finished goods that will inevitably result from continuous factory output.[13] This strategy is the diametric opposite to those of innovative factories from VW, Porsche, and Mercedes.[14] Virtual, or "solution" factories[15] do more than focusing on flexible response to volatile demand. They go so far as to adapt entire manufacturing processes and associated employee behavior to mobilize the manufacturing function into a customer-centric team. Porsche, for example, dedicates a single worker to the entire production process of a car. HP's server (dubbed Solution) factory differentiates tasks by competence level and uses annualized labor contracts that permit flexibility as precise as one hour, and this in labor markets as intransigent as Germany's.

Myths and realities

Arguably, the most important task of the Innovation Diffusion group at HP is to question the assumptions of managers who ask us to support new programs and projects. In the course of our work, reviewing plans to launch anything from the latest IT solution, to a new logistics or process improvement program, it becomes apparent that a number of myths drive management practice. For all its self-imposed analytical rigor and relentless publication, academic research has done little to make ideas a foundation for feasible action for the very community that is their laboratory. Here is a less-than-exhaustive list of influential, thus dangerous myths that we continue to encounter.

Real-time information. There has been an ongoing craze for real-time information, partly driven by the deployment of powerful information systems and as well as availability of Internet-based information technologies. The reality of most supply chains is, however, that supplier lead times far exceed the cycle times of assembly and test, as well as the typical delivery lead times. These disparate time scales can lead to distorted views on business performance; even worse, they can drive uncoordinated decision-making.[16]

As anyone will recall from their Physics 101 lab, sample rates and signal frequency can predetermine or impair our perspective on what is happening. Therefore, when the data entry frequency is erratic (causing us to miss signals), or if the samples are based upon different units of measure, or if they come to view discrete and arbitrary time windows (while we expect them to be continuous and systematic), the data sampling will distort the analysis of actual system behavior.[17]

In daily business operations, the data points are typically sparse, and selectively consolidated to reflect quarterly and annual financial performance-hardly the normed laboratory for sound statistical analysis. This illustrates that the striving for real-time information, (presumably in the hope of zooming in on reality faster), in order to improve business results is often misguided. It also begs the question: "How can we exploit this elusive reality to guide business decisions?"

The tool of choice for many businesses eager for real-time insight, uncritically implemented, then baffled by the disappointed return on investment (ROI)[18] is the data warehouse.[19] Well-designed systems or analytic projects fail to bring the expected returns (and the expectations upon technology remain astonishingly high, emotional, and lacking qualification in functionality) because of the most banal of reasons. The operator is a minimum-wage worker. Information that must be shared is parked because collaboration between workers is not rewarded and employees benefit more from keeping knowledge to themselves. Information quality depends not upon whether it is in "real-time" or not, but how appropriately it has been structured for search and query. Are the key fields that the system developers anticipated relevant to the question? Did the system designers account for the possibility of commercial airlines flying into high-rises?[20] The training, quality awareness, and availability of data entry operators, all of which will cost money, have another profound impact, which is rarely factored into academic correlations (or project budgets for that matter.) As Curt Hall regretfully concludes "Many corporate data warehouses fail, not because they do not have the correct information, or "enough" information, but because their end users simply do not know how to use the warehouse or how to apply their findings."[21]

Successful examples of balancing supply chain dynamics with decision-making horizons can, however, be found at HP. For example, the Automated Inventory Replenishment (AIR) program, which replenishes channel inventory, uses a weekly review period. IT systems at HP's server or "solution" factory in Gueltstein align inventory allocation with the subsequent manufacturing process, with the final shipments, and most critically, with the delivery times expected by the customer.

Vertical integration. With the advent of the PC, and IBM's decision to "outsource" the microprocessor, operating system, and firmware, vertical integration is a passed station in the developmental journey of the high-tech industry. Most products include a host of third party hardware and software, creating a network of companies contributing to the customer value. This approach has spread to other industries that are often perceived to be "vertically integrated". An example is steel, where so called mini-mills process steel sourced from global suppliers, rather than create steel from scratch.[22] When analysts or researchers call for vertical integration they ignore the realities of value collaboration networks. Even if the information flow were to be successfully integrated, the metrics driving the performance and behavior of the individual companies would deter operations as a vertically-integrated supply chain. Moreover, increased specialization in both competencies and assets, makes it virtually impossible to integrate from a return-on-investment perspective. As the LSI Logic example (embedded within a consortium to build capacity), illustrates, the interdependencies of the players, their disparate objectives, and diverging metrics make vertical integration a myth. Once you have lost the financial and human capital base for a vertical integration, it is very difficult to rebuild that foundation, let alone become a competitive player. Moreover, the trends of shedding fixed assets and managing brands have created an economic landscape that favors risk mitigation through specialization. Finally, because so many players in the high-tech industry act as suppliers or partners of their own competitors, real incentives exist to maintain sub-optimized processes and perpetuate self-interested behavior.

Outsourcing reduces cost and creates flexibility It is striking how, in blatant disregard of the known trade-offs, outsourcing is being decreed as a blanket supply chain management strategy in the high-tech industry. Any cost reductions from outsourcing are typically the result of a myopic focus, abbreviated time horizons, or a zooming-in on one isolated driver of supply chain costs (like cost-per-part). Reviewing the whole picture, however will likely arrive at entirely different conclusions about real benefits. Neglecting the impact of logistics, escalation, rework, warranty, support, and cash-flow is typical of a hasty and myopic view of supply chain costs. Another obscuring element is control and visibility.[23] The net result of outsourcing is an accumulation of hand-offs in a supply chain, which inevitably increases delays, and opens up the abyss of misalignment. Reduced visibility decreases overall flexibility and responsiveness. It can also offset the flexible use of capacity enabled by the sharing a supplier with other customers. Furthermore, how do you incent an autonomous supplier to meet your particular business needs? The OEM will be confronted with increases in component prices in close to real-time, whereas drops in cost will understandably be kept well-concealed from the OEM. HP's former chief procurement officer Corey Billington concludes in a recent interview "It is a characteristic of outsourcing. ... Companies outsource functions that previously were not well-priced to begin with." Outsourcing creates an advantage, but then you have hidden costs and inflation in loopholes that company employees can't negotiate away.[24] The intrinsic dependency of an in-house operation allowed control over metrics and information, but this visibility becomes elusive in the outsourced configuration. An industry of auditors, contract and systems designers may well mushroom to address the multimillion dollar loopholes which open up in poorly-regulated divisions of labor. The charges of their professional services will have to be tallied against any savings made by contracting away the specialized function.[25]

If we take the lesson of component pricing to heart, we will realize that the drive to remove assets from the books does not automatically yield sound business results, but can actually aggravate a weakness. A worst-case scenario could involve transferring ownership of a manufacturing operation without taking on additional users of this capacity load. This half-hearted move cancels out the main objective for outsourcing i.e. risk-pooling of capacity, actually increasing the cost of overtime, or destroying the viability of the assets. Finally, once outsourced, it is expensive and difficult to reestablish a competency. In some European countries like Great Britain it is even illegal to rehire comparable job positions after shedding the headcount.[26] Tax and duty advantages could also be lost in the bargain, creating anything but the hoped-for increase in flexibility and reduction of cost.

Build-to-order means velocity and flexibility. Building to customer order (BTO) is typically in conflict with reducing costs. It is also an over-simplification: while it takes minutes to assemble a PC (and additional time to burn-in and test), the longest supplier lead times are in the order of months. This time discrepancy requires sophisticated forecasting and planning to ensure that all parts are available. This is complicated by the fact that, in the high-tech industry, the longest lead times may be longer than the product lifecycle! The inability to accurately forecast is often offset by using supplier hubs, where OEMs do not have liabilities for the parts until they are pulled from pre-delivered inventory. For generic parts with diversified demand patterns, suppliers can probably afford this. However, unique or custom-made parts which cannot easily be sold on the remaining market will drive up costs for the supplier, which when written-off will either negatively impact their business health and viability, or else be passed on as generally increased prices to the manufacturer. Thus, although the vendor delivering to pooled hubs may be apparently responsive and flexible, the end-to-end supply chain, typically, is not. Only deliberate holistic decisions about where to hold inventory, how to share both risk and benefit can truly increase velocity and flexibility. Disregarding overall supply chain performance by punishing suppliers does not. If the supply chain cash flow is taken into account, this picture becomes more bleak. That is, payment terms reduce the cash available to the upstream players, diminishing their ability to invest in flexible capacity, or in the inventory required to meet end-user demand. Therefore, in markets where time-to-revenue is critical to business success, production ramp-rates might be artificially constrained by the cash position of suppliers. Without assurance of component supply for the BTO operation, it is an illusion to expect that your operation can instantly meet volatile customer demand. Furthermore, increased demand volatility requires larger inventories of parts, making inventory management a game of survival. In high-tech manufacturing, products devalue incessantly over their lifecycles, therefore pulling component inventory on-demand is advantageous, because it lets you exploit market prices to either increase margins, or lower price in order to drive demand. When part prices are increasing (recall the skyrocketing DRAM market of 2000), this may not be as beneficial. These factors also determine the effectiveness of auctions as a sourcing tool for any business case, since locking-in a price can induce a runaway competitive disadvantage, capped only by the volume of your spend.

Modeling reduces costs. Modeling can be an instigator for adopting better processes or organizations, but in itself does not change behavior. Involvement of all stakeholders-the people that can make or break a project -is the prerequisite to successful implementation. Otherwise, experience has shown that the lone modeler exposes herself to challenges from all sides to the legitimacy of the proposed model and its conclusions: the data, the assumptions, the approximations, the results, and how to apply its outcome. Ironically, these battles of authority typically drive up cost, rather than saving anything. Furthermore, models rarely take into account the return-on-investments, or any business practices that could thwart the projected return. A famous example is the application of postponement to the manufacturing process of HP's DeskJet printers.[27] Although the theoretical model initially predicted savings in FG inventory, the actual savings were significantly higher when the less conspicuous packaging function discovered its opportunity to reduce cost.[28] This reiterates the need for cross-functional involvement, and the positioning of models as an enabler of decision-making, not its driver.

"Best in class" creates competitive advantage. History suggests otherwise, most successes in technology adoption can be linked to the commercial exploitation of technologically mediocre products (like video, PCs, combustion engines, or filament light bulbs). Marketing, timing, and distribution networks seem, instead, to be critical to success. Moreover, dependent upon your industry, any competitive advantage of "best-in-class" may have a very limited longevity compared to development costs, disappointing expectations of a good return-on-investment. Settling for "good-enough" product developments and focusing on agility in your supply chain are more typical characteristics of success. HP's home PCs, for example, became successful in the retail market when they focused on processes. Information technology was adopted pragmatically to fit the needs of these processes, and geared up to support a volatile, customer-driven commodity marketplace. Its success can be attributed to a can-do attitude of the team, rather than adoption of best-in-class technology. Revering benchmarks can create a copycat (necessarily a follower) attitude to business management. As the example of unforgiving consumer markets demonstrates, competitive advantage will require courage to try the untested, and bold innovation.

Killer application. Media and companies alike uphold the myth that the success or failure of a technology is linked to a killer application. In fact, technology adoption is much more an outcome of customer attitudes[29] or legislative force (for example for seatbelts and airbags). Furthermore, if the application lacks standards, customers may be wary or become frustrated. A case in point is the of cell phone technology infrastructure in the US, whose multiple standards have deterred widespread adoption of the short messaging services that are popular in Europe or Asia.[30] Similarly, the adoption of broadband Internet connectivity seems to be much more related to the perceived trade-off between price and performance than to any lack of content, or the capabilities of web browsers.

The same mechanism also drives transforming software applications innovation. Software companies use industry leaders as testing ground for "best practices". In the current economic landscape, few companies are willing to invest in software where the survival likelihood of the supplier is perceived questionable. This decreases the likelihood of innovation or adoption of break-through technology, both from a testing and a purchase perspective.

What does this mean for Operations Research?

The field of operations research continues to refine its methods of modeling and programming, oblivious to the fact that the problems they presume to address may be inappropriately captured and set up. Mathematically indisputable, the outcome of these computations may not only be irrelevant, in the worst case, it may be dangerously misleading. The impact of false assumptions on our ever-changing environment makes the need to (re-)validate well-established operations research practices (like economic order quantities) all the more urgent. An often-used approximation, large numbers, must be challenged in the face of contemporary industry trends. Solutions with low recurrence rates (because they have unique configurations), make it very hard to use large product volume approximations. This will come as no surprise for those working on processes like store-level replenishment. In retail operations, decisions on one, two, or three inventory units can have profound implications on cost, sales, and credit.

Similarly, since partial shipments tend to be unacceptable to customers, the relevance of typical fill rate optimizations become questionable. The cost of missing a shipment, resulting in rework cost, penalties, and missed sales, might well exceed any benefits derived from conventional replenishment policies.

Finally, even with apparently pervasive information systems, research must face up to the paradox of inadequate data in real business environments. The data that exists in most organizations is often incomplete, sampling rates are low, or data entry by minimum-wage, temporary staff has occurred at irregular frequencies. These characteristics may well violate the input requirements of algorithms.

For the field of operations research, incessant change and inescapable dependencies imply that certain widespread models and algorithms must be revised. The models, which rely upon single node approximations, assumptions of independent demand, and the review of material, information, and finance flows irrespective of their relative impact on each other, do not necessarily reflect current business realities. Most importantly, if applied in the wrong context, these models might end up actually increasing inventory cost, or diminishing availability.

Conclusion: Cases of successful collaboration

Operations Research is the historical foundation of Supply Chain Management and has lost neither its impact nor its relevance, in spite of the dramatic shift in what we claim to be problem definition. Single-node approximations have accompanied single-criterion decision-making in the shop floors and offices of our industry. Both interested parties-effectively building a dichotomy of the researcher and the researched, subject and object-would benefit from an explicit definition of what they do not know. We can imagine a day on which a GM will intuitively speak of "pushing down the efficient frontier" instead of "reducing inventories", or the OR professor will enjoin a class to "build a model, then pilot, then revise your model in a continuous feedback loop,..." instead of simply "build a model". Revising the prevalent, uncritical attitude that models "should yield an answer" would help too. Let us never forget that models are built upon highly specific assumptions. When their results are applied back into practice, the current snapshot of reality may have little to do with the original problem definition. Moreover, unquestioning faith in scientific models, presumably in the hope of imposing rigor and meaning upon chaotic systems, is an attitude which percolates right down to industry applications. How often do users delegate thinking, trusting hidden algorithms and their hard-wired parameters, turning to the black box to authoritatively provide "the answer". Ideally, however, a deliberately-selected simulation tool should enable decision-makers to reveal high-impact uncertainties and dependencies in the input scenario. Following this evaluation, the testing and piloting of an implemented solution should never, ever be skipped.

Clearly, the metrics and life-cycle of innovation in academic research and industrial practice differ. Several months or even years is affordable (and respectable) for the in-depth pursuit of an idea at a university, whereas a single quarter in industry can decide the fate of an inventory management strategy. The sample rates of academic case studies would profit from velocity, i.e., more frequent and continuous interaction with the managers in industry. At present, however, as long as internal standards of rigor are met, validating the assumptions and delineating the operative scope of a model do not even make it onto the list of academic priorities. As our cases illustrate, the isolation that results from addressing a narrowly-defined audience can produce false confidence in the readers who were implicitly excluded, with fateful consequences in any subsequent onsite implementation.

With a culture rooted in Bill and Dave's ties to Stanford University, HP has been fortunate enough to look back upon a history of fruitful exchange with academic institutions. In our observations, the success of HP's collaboration with research institutions depends upon simple mechanisms, and can take on a variety of forms. Professor Eric Johnson from Dartmouth's Tuck School works every summer upon projects within the company, removed from his university's constraints and scholarly detachment. Johnson also co-teaches our team's basic supply chain management course to HP staff and managers around the globe, receiving immediate feedback on the feasibility of theory, as well as what is developing at the industry's cutting edge. Members of our Innovation Diffusion team regularly run the beer game for Professor Warren Hausmann's students at Stanford university, affording them direct insight into how current corporate operations are enacted by the game.

No one company can afford to reflect and analyze a phenomenon with the breadth and sustained concentration of a scholar. Not only is time a factor, but market conditions prohibit the objective distance that serious analysis demands. Therefore, one of the most valuable forms of partnership is the benchmarking process, in which the academic institution plays the neutral moderator, assessor, and filter. Professor Arnd Huchzermeier of the Wissenschaftliche Hochschule für Unternehmensführung in Vallendar in Germany, and Professor Ludo Van der Heyden of INSEAD collaborate with popular European business magazines to annually honor what they call the "Best Factory".[31] This university-sponsored "contest" overcomes barriers of trust and communication to provide value to all the parties involved.

The scholars gain access to a wealth of cutting-edge manufacturing innovations by directly polling and interviewing a sample[32] of companies. At the ensuing plant tours, they encounter an otherwise uncharacteristic willingness of their (delighted) subjects – including key executives – to share details. The business press publishes a compelling story, legitimated by a prestigious academic institute, complete with interviews of top executives and exclusive photos. The laureates of the Best Factory benefit from the free and glowing marketing in the magazine's cover story. The competition concludes with a conference-cum-awards ceremony, at which each laureate presents the specifics of their plant, followed by commentary and contextualization by the panel of professors and journalists. Not only do the practitioners from industry gain valuable benchmarking and market insight, the neutral ground inevitably becomes a forum for networking and informal exchange.

The TECTEM (Transferzentrum für Technologiemanagement) benchmarking institute at the University of St. Gallen puts the same principles into practice to select and honor best practices in many other areas of operations management. A dedicated research team from TECTEM interviews and visits European companies[33] on an ongoing basis. Moderated by the university, benchmarking participants are invited to present themselves to one another on-site, and are provided with detailed reports of each event. Professor Daniel Corsten's recommendation at a European Supply Chain Management conference[34] reflects what HP has successfully practiced for years: if manufacturing companies maintain R&D teams to develop products, why not create R&D initiatives around internal processes? HP's supply chain think tanks, and innovation diffusion teams to carry out internal research and development of their process environment.[35] As the preceding catalogue of myths makes abundantly clear, questioning the assumptions of project managers and program sponsors is one of our most critical tasks – and has deterred many a doomed investment.

Overall, HP's successful encounters with academic research can be traced to individual passion, uncompromising ethics, and heroism. Interestingly enough, we are aware of no instance of profound and sustained collaboration that was the outcome of an institutional program. Creating a partnership cannot be decreed: trust is neither goodwill nor even feeling good, but built upon a balance of power, and the implementation of a win-win charter. The publications that defined Supply Chain Management for a generation of managers and scholars established both Professor Hau Lee and Dr. Corey Billington as experts.[36] Both sides benefited from telling the story: HP consolidated its brand as a manufacturing innovator[37] and Professor Lee's academic reputation was anchored. Their mutual esteem and prodigious personal energy were key to fruition.

The details of ethical behavior are banal: sharing the credit for a successful innovation in industry makes the pie you are dividing up bigger, whereby a scrupulous respect for intellectual property does not hurt. Let us articulate another truism: authorship implies authority, as both the source and owner of invention, a sensitive issue for many project managers (who may not qualify academically through a list of publications per se), who understandably expect recognition in any article based upon their labor and courage. Need we say that the prospect of being reduced to scientific material by a publishing scholar, divested of the active (thought) leadership that led to business success in the first place, will become a barrier to communication and trust in the future.

When HP looks back upon its history of collaboration with universities, it has been these fine details that enabled (or sabotaged), success. Let us continue to combine forces with professional maturity. It is not at a safe distance, but at an unaccustomed proximity, that universities and businesses will more effectively address the burning platforms which can threaten the existence of whole industries.

Notes

[1] In a recent article, the authors come to the conclusion that "In truth, supply chain management (SCM) remains more of a pipe dream than a reality." In the article, unfortunately, implementing SCM is stated to be the equivalent of [launching] "a number of initiatives, including efficient consumer response (ECR); vendor-managed inventory (VMI); and collaborative planning, forecasting, and replenishment (CPFR)." The authors proceed to detail some of the barriers and propose actions to overcome the obstacles, some which will require longer time horizons than the business cycle will permit (like "Develop a new breed of manager"). The list of barriers is made without analysis of the ("best") practices which construct them, i.e. no comment is made on the risks of blind faith in one-size-fits all approaches. Although no sick person would mindlessly swallow a pill that once cured a distant relative, common-sense habits of diagnosis and measuring dosage are often lost in the business rush. To their credit, the authors' concluding enjoinder is to "Engage in More Practical and Applied Research!" Moberg, C., Speh, T., and Freese, T. (2003). "SCM: Making the Vision a Reality", *Supply Chain Management Review,* September/October, p34–39.

[2] Decock, B., Page, M., and Flebut, J. (2002), "How to generate a business advantage in a fiercely competitive manufacturing environment," *HP Webcast Series,* August 22, `http://www.on24.com/clients/ hp/index.asp?sessionid=14:` The factory will in fact be built by a consortium of suppliers. The role of the manufacturing company in the "center" of the supply chain as a determining agent of its structure and flows has been transformed by this innovation of collaborative ownership.

[3] Moberg, C., Speh, T., and Freese, T. (2003). "SCM: Making the Vision a Reality", *Supply Chain Management Review,* September/October, p34-39. The article gave a checklist of factors (many of which require sweeping change with years of costly implementation), for successful implementations, while perpetuating uncritical trust in turnkey initiatives i.e. suggesting that Supply Chain Management will produce measurable business results if only all these actions were exhaustively carried out. The sometimes banal habits of successful supply chain management (identifying your industry's characteristics, network mapping, locating bullwhip triggers, prioritizing business-critical metrics, identifying key stakeholders, then communicating diagnostic results to get buy-in, etc.), do not seem worth mentioning.

[4] We do not recommend a surge in professorial consulting start-ups (although some experience of the consequences of academic analysis will benefit research, without doubt).

[5] This is clearly one of the reasons why formulaic, checklist, one-size-fits-all SCM initiatives like VMI, CRM (customer relationship management), BTO (build-to-order), etc. cannot reliably work unless they have been identified as the "pill" which should cure the diagnosed illness, within the known boundary conditions.

[6] That inventory is always a liability because it is nothing more than superfluously locked capital is an alarming myth that we battle constantly. It is the rare manager who declares a strategy "to push back the efficiency frontier", rather than simplistically "reduce inventory" on her P & L.

[7] Lost market share in the high-tech industry when the market shifts is rarely recouped. Who remembers for example the first "portable" computer manufacturer Osborne (See for example Billington, C., Lee, H., and Tang, C. (1998). "Successful strategies for product rollovers" *Sloan Management Review,* Spring, 39(3):23–30.).

[8] See for example A.C. Marquez and C. Blanchar's presentation at INFORMS, Salt Lake City, 2000 entitled "Evaluating Marketing Strategies using System Dynamics Models." Four subsequent HP projects using improved modeling techniques illustrated this phenomenon, in different supply chains, operating in different channels.

[9] 1999, HP internal paper

[10] Currently a similar capital issue is hampering investment and capacity allocation in the shipping industry where a "pigs cycle" and the "bullwhip effect" are creating delays, distortion, and amplification of responses to demand dynamics: Matthews, R.G. (2003). "A surge in ocean-shipping rates could increase consumer prices", *The Wall Street Journal,* November 4.

[11] Metrics often create undesired behavior. In successive collaborative projects with retailers HP encountered again and again that aligning trucking activities with dock capacity yielded significant costs savings in inventory while improving availability. This is a direct result of the performance metrics of truckers, who are measured on miles driven, not time. This decreases trucking capacity utilization and drives up supply chain costs. Interestingly, industry sees changes to these metrics as threatening, clearly lacking a holistic view of supply chain costs. Machalaba, D. (2003). "Cost of trucking seen rising under new safety rules", *The Wall Street Journal,* November 12.

[12] "Strategic Cost Management in the Supply Chain: A Purchasing and Supply Management Perspective," written by Lisa M. Ellram, a professor of supply management, Bebbling professor of business, Arizona State University, available through CAPS research. See also Ellram, L.M. (2003). "A Prescriptive Model for Cost Management in the Supply Chain," *ASCET,* v5, July 7, `http://www.ascet.com/documents.asp?d_ID=1985.`

[13] White, G.L. and Lundegaard, K. (2002). "U.S. auto sales accelerated 13%, Driven by Deals," *The Wall Street Journal,* September 5, pA1.

[14] See for example Anonymous (1999). "Porsche to make new sports utility vehicle in Leipzig," *Automotive Intelligence News,* November 16, p2; Mudd, T. (2000). "Back In High Gear", *Industry Week,* February 21; Smith, T. (2001). "VW factories in Brazil look for export markets", *The New York Times* September 3, pC1.

[15] Kuper, A., Kahn, D., Schmid, H., and Thakur-Weigold, B. (2002) "Delivering solutions to customers: high-velocity manufacturing," *ASCET,* v4, May 16, `www.ascet.com/documents.asp?d_ID=998.`

[16] Forrester, J.W. (1958). "Industrial dynamics–a major breakthrough for decision makers," *Harvard Business Review,* v36, n4, p37–66. Forrester, J.W. (1971); *Principles of systems,* Wright-Allen Press. See also `http://sysdyn.clexchange.org/sdep/papers/D-4224-4.pdf;` Lee, H. Padmanabhan, V., and S. Whang, (1997), (1997). "The Bullwhip Effect in supply chains," *Sloan Management Review,* Spring, pp93–102; Kuper, A. (2000). "Managing supply chains while moving at Internet speed," *Cutter IT Journal,* March, v13, n3, p17–22.

[17] See for example Shearer, Murphy, Richardson, *Introduction to system dynamics,* Addison-Wesley, 1971; chapter 9: Interference, of Hecht, E. and Zajac, A., *Optics,* Addison-Wesley, 1979; Wilson, J., Hawkes, J.F.B., *Optoelectronics: an introduction,* p 297, van der Meulen, S.F., *Fysische Meettechniek I* and *Fysische Meettechniek II,* Technische Hogeschool Twente (University of Twente), 1981, first-year applied physics university course reader; Bendat, J.S., Piersol, A.G., *Random data analysis and measurement procedures,* New York, 1971; F.R. Connor, *Modulation,* E. Arnold, London, 1973; K. Küpfmüller, *Einführung in die theoretische Elektrotechnik,* Springer, Berlin, 10. Aufl, 1973, chapter 5; Elgerd, O.I., *Control systems theory,* Mc Graw-Hill, New York (1967); an understanding of Fourier or Laplace transforms provides a mathematical basis.

[18] Although half of all organizations rate their data mining efforts as moderately successful, just 6% believe their data mining efforts have been a“major success.” Probably the most startling finding is that 44% of organizations believe that their data mining efforts either have “not contributed much in the way of success,” or have “not contributed any real value to their organization’s business efforts.” In other words, for 44% of organizations, it appears that there has been little or no business advantage gained from their data mining efforts to date.”

[19] Hall, C. (2002). “How companies rate the success of their data mining efforts,” *Cutter Consortium,* May 8. As Curt Hall concludes “A data warehouse is only one part of a decision support solution.”

[20] Hall, C. (2002). “The terrorist information and prevention system” *Cutter IT Consortium,* July 23.

[21] Ibid.

[22] See for example King, Jr. N. and Guy, R. (2002). “So far, steel tariffs do little of what President Bush envisioned”, *The Wall Street Journal,* September 13, pA1.

[23] See Sullivan, L. (2003). “Outsourcing’s Hidden Costs,” *Electronic Buyers’ News,* www.ebnonline.com, May 10. Laurie Sullivan summarizes “Electronic manufacturers embracing the outsourcing trend are incurring hidden costs because of sloppily written contracts and a lack of resources to audit component prices and services.”

[24] See also, Billington, C. and Kuper, A. (2003). “Trends In Procurement: A Perspective,” *ASCET V,* Montgomery Research, San Francisco, p102–105.

[25] “... OEMs approach these services like it will solve all their problems ... Companies think they can turn over the function to a consulting firm that will find all these holes, and that is not always the case. If as a company, you are not keeping track of what the market is doing and auditing the audit company, you will still lose millions.” Ibid.

[26] See for example Atkinson, A. (2001). “Outsourcing manufacturing: creating a value collaboration network”, *HP Webcast Series,* February 15, http://www.on24.com/clients/hp/index.asp?sessionid=40.

[27] See H. Lee, C. Billington, and Carter (1993). “Hewlett-Packard gains control of inventory and service through design for localization”, *Interfaces,* v23, n4, pp1-11.

[28] Howard, K. (2001). “Beyond postponement: regional logistics effectiveness,” *HP Webcast Series,* March 21, http://www.on24.com/clients/hp/index.asp?sessionid=43 for a more recent perspective see www.hp.com/go/manufacturing.

[29] See for example Piszczalski, M. (2002). “Five myths of telematics.” In *Automotive design and production,* http://www.autofieldguide.com/columns/martin/0702it.html, July, as well as Zetie, C. (2003). “The myth of messaging,” *Information Week,* February 17, and Jasper, J. (2002). “The Quest for the Killer App”, *supportindustry.com,* February.

[30] See for example Labarge, R. (2002). “Can your clients play your DVDs,” *Digital Video,* July, pp20-32.

[31] For full details of the Best Factory Award along with case studies of past laureates, see Loch, C.H., Van der Heyden, L. Van Wassenhove, L. Huchzermeier, A. and Escalle, C. (2003). *Industrial Excellence: Management Quality in Manufacturing.* Springer-Verlag, Berlin.

[32] Due to the voluntary application process, the sample is self-selected and cannot be exhaustive. The magazines invite companies to download a questionnaire and submit it to the selection committee. In spite of the questionable statistical significance of the judgment of who is "best", it provides enormous value to both industry and scholarship.

[33] The sample is selected by the university. The response rate to the invitation results in a form of self-selection.

[34] Professor Daniel Corsten's concluding remarks at the "Supply Chain Management" conference organized by the University of St. Gallen's Institute for Technology Management, January 29-30, 2002 in Zürich, Switzerland.

[35] See Kuper, A., and Branvold, D. (2000). "Innovation Diffusion at Hewlett-Packard," In Johnson, M.E., Pyke, D.F., editors, *Supply chain management: innovations for education, Production and Operations Management Society (POMS) series in technology and operations management* v2, pp205–315.

[36] Lee, H. and Billington, C. (1992). "Managing supply chain inventory: pitfalls and opportunities," *Sloan Management Review,* v33, p65–73; Lee, H. and Billington, C. (1995). "The evolution of supply chain management: models and practice at Hewlett-Packard," *Interfaces,* v25, n5, p42–63.

[37] Turned into a successful marketing program by its Manufacturing Industries team, which sells thought leadership to external customers as a differentiator from those consulting firms which will never have to implement their designs themselves.

This page intentionally left blank

Chapter 13

SUPPLY CHAIN MANAGEMENT: INTERLINKING MULTIPLE RESEARCH STREAMS

James C. Hershauer
W.P. Carey School of Business
Arizona State University

Kenneth D. Walsh
Department of Civil and Environmental Engineering
San Diego State University

Iris D. Tommelein
Civil & Environmental Engineering Department
University of California at Berkeley

Abstract This chapter represents a view of the evolution of the supply chain literature from a wide range of perspectives including operations management, logistics, purchasing, and information technologies. First, a chronological view of the field is presented. Next, eleven major research streams are summarized.

The streams are:
(1) inventories,
(2) global supply logistics,
(3) buyer/supplier dyads,
(4) communication and the Internet,
(5) lean supply chains,
(6) process analysis,
(7) power,
(8) mass customization,
(9) alliances,
(10) market structures, and

(11) environmental life cycle.

Conceptual models are provided to depict the influence of industry. Diagrams depicting interfaces among stakeholders in the aircraft, marine, and construction industries provide examples of more complex relationships as contrasted with the typical linear models of a retail industry. In summarizing similarities and differences among research streams, primary and secondary emphasis on seven areas of research are provided for each of the streams. Finally, four categories of organizational supply chain maturity are suggested. Progressive maturity levels are pretender, follower, thinker, and industry leader.

1. Historical Perspective

Looking back at the evolution of production and operations management as a field, one can see a progression through several phases:

1 (1900-1949) The formation of basic foundations during the first fifty years;

2 (1950-1969) An explosion of operations analysis tools and models during the next twenty years fostered by rapid technological change;

3 (1970-1979) Followed quickly by a ten-year focus on productivity fostered by a new global awareness;

4 (1980-1989) A ten-year focus on quality fostered by awareness of the customer and the growth of the global marketplace; and

5 (1990-1999) A ten-year focus on speed and flexibility fostered by an emphasis on the customer and use of information and communications technology.

6 (2000-present) An on-going focus on supply chains fostered by an era of global operations and competition facilitated by Internet communications and Web transactions.

Henry Ford was probably the first modern-day manager to recognize the immense benefits of coordinating all subsystems in the flow of all raw materials to a final consumer product. Of course, he had no choice: he had to develop an integrated supply chain out of necessity. Supplier organizations were yet to develop in order to take advantage of specialization, but fragmenting the supply chain at the same time.

Ford's approach was to own and control all manufacturing and assembly.

> "Ford ... vertically integrated to manage the entire supply chain within his own organization ... employees mined iron ore in the firm's own pits. Ford ships and tractors transported the ore ... Ford cranes unloaded it. Ford steel mills processed the iron ore to make steel plate from which Ford factories built the Model T ... Inventory never sat unused for long. Ford boasted that iron ore unloaded at his River Rouge plant became steel components in a Ford truck rolling off the assembly line within 48 hours." (Melnyk and Swink, 2002)

Now that sounds like a supply chain with minimal waste of time, inventories and motion!! And you could have any vehicle that you wanted as long as it was identical to all the others and black! Ford was trying to reach the masses with a single standard vehicle designed with the sole objective of cheap personal transportation. He also chose to allow gas station operators to sell the vehicles for him. These small local operators could then provide both the gasoline and service to keep the cars running. In the United States, many former gas station owner families were eventually granted franchisee status and these grew into the current powerful dealership network.

Alfred P. Sloan, Jr. (1963) at General Motors then set out to determine and codify how to make Ford's approach managerially possible in a multiple-product environment. More complex supply chains were created. From 1950 through 1969, pioneers like David W. Miller and Martin K. Starr (1960), C. West Churchman (1961), and Jay W. Forrester (1961, 1969, 1971) developed the systems-level models for decisions in more complex supply chains. Forrester (1961) was one of the first to examine integrated supply chains in his book on Industrial Dynamics. From about 1970 through 1999, functional specialists focused on specific decisions within operations, logistics, and purchasing by considering the elements of value in order of cost, quality, flexibility, and then time.

Supply chain management came into being as a field of its own as the realization grew that real performance improvements could be obtained by systemic modifications of the traditional disciplines of (1) operations management, (2) logistics, and (3) purchasing. Conceptually, such an approach had been widely understood for many years, but the evolution of distributed, inexpensive information technology resources (Evans and Wurster 2000) has made it possible to share information at unprecedented levels. Each of these foundational disciplines, of course, has a particular perspective. Operations management has traditionally focused on resource management, scheduling, quality, and inventories. Logistics has traditionally focused on customer service through managing the location and transportation of materials. Purchasing has tra-

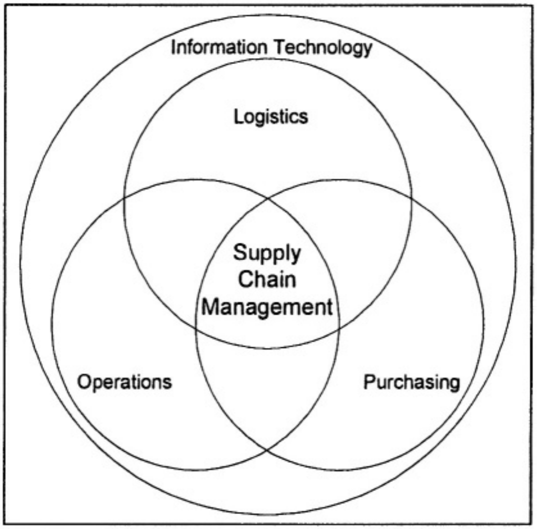


Figure 13.1. Supply Chain Management as an Interlinking of Traditional Disciplines.

ditionally focused on obtaining materials, supplies and services and on managing relations and contracts with suppliers. Information technology has traditionally focused on tracking and recording transactions, storing and retrieving data, and creating information for decision support.

While the science of supply chain management arises largely from interlinking the underlying traditional disciplines, simply integrating these four views only tells part of the SCM story. Reading the supply chain literature, one is faced with a vast number of papers, books, and monographs, each of which reflects the perspective of its authors. In this chapter, we will categorize the most influential research streams that underlie the approaches to supply chain problems. While we have cited Web sites and literature that we think will help one start a reading of the relevant literature, we know we cannot be comprehensive in our citations. There are thousands of related references. We hope that you will find the few that we have chosen as helpful as we have in grasping the different interlinking research streams. We wrote this summary to assist the reader in understanding the literature, and in adapting solutions presented to the circumstances of their own particular supply chain interests.

Within operations management, logistics, and purchasing, different schools of thought co-exist regarding the central concepts of most import, and echoes of these approaches can be heard in the supply chain

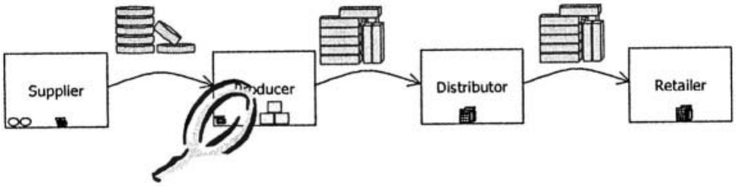


Figure 13.2. Inventory View: Supply Chain as Collection of Related Inventories, Inventory Levels and Policies Receive Special Attention.

literature as well. In addition, the emerging field of supply chain management has facilitated the rise of new schools of thought more broadly cast across the boundaries between the foundational disciplines. Because information technology is such an important enabling technology, perspectives from the peculiar technical problems with information technology deployment also arise. The traditional research streams are summarized first and then the related concepts of lean supply chains, process analysis, power in chains, mass customization, alliances, market structures, and environmental life cycles are reviewed.

2. Research Streams

2.1 Inventories

In many ways, a good deal of the improvement in supply chain performance observed today can be attributed directly or indirectly to the reduction of inventory levels system-wide. The benefits of such an approach are obvious: less capital at risk, lowered risk of shrinkage, and reduced inventory management expenses, just to name a few. The potential costs are risks such as stock-outs and lack of order fulfillment. Since inventory reduction strategies have been so effective, it is tempting to view supply chains as a collection of related inventories. With this view, the analysis of a supply chain focuses on inventory location and optimization of quantities and tradeoffs to achieve desired supply chain delivery and cost. Although Figure 13.2 shows a magnifying glass on producer inventory, the idea is that inventory at all locations in the supply chain are put under the magnifying glass of extensive cost analysis. The six types of supply chain inventory represented in the figure are (1) cycle inventory, (2) safety stock inventory, (3) market inventory, (4) pipeline or systemic inventory, (5) anticipation inventory, and (6) coordination inventory.

Cycle inventory analysis evaluates the tradeoffs between lower order and setup costs with larger batches and lower inventory costs with smaller batches. Safety stock analysis evaluates the tradeoffs between the cost of holding excess inventory versus the cost of stockouts. Market inventory exists to exhibit and promote product and to meet demand immediately. Market inventory costs are traded off against costs of lost sales and lost market share. For example, even though information technology and manufacturing technology used in the automotive supply chain would permit almost complete elimination of the huge market inventories at dealerships, most customers are not willing to wait only the few weeks now needed to deliver an ordered vehicle. Pipeline inventory is generally computed as the average daily demand times the total time in the pipeline. Although the term pipeline has frequently been used historically to describe the inventory in transit in the distribution and supply channels, it can also be used to analyze the systemic inventory present in the entire supply pipeline. Anticipation inventory refers to any inventory held in anticipation of possible special interventions in the flow of supplies. Special interventions could include such events as plant shutdowns, strikes at suppliers or transportation providers, weather related delays, and demand surges. Coordination inventory exists to avoid other costs involved in scheduling the use of multiple resources. For example, construction sites normally have extensive inventory in lay down areas to avoid the costs of idle craft labor and construction equipment. Also, if a materials requirement system experiences a late delivery of one part, other parts will be carried in inventory until the late part is delivered.

Historically, the emphasis has been on minimizing cycle, safety, anticipation, and coordination inventory. A supply chain perspective changes the focus to eliminating market inventory and drastically reducing pipeline inventory. For example, reducing the total supply chain length from 500 to 50 days reduces pipeline inventory by 90% (assuming inventory is distributed evenly in the supply chain) and makes an obvious reduction in response time. If inventory is not evenly distributed, the reduction may be slightly less. Eliminating market inventory completely by direct shipment from the first-tier supplier may revolutionize an entire supply chain (personal computers for example).

2.2 Global Logistics and Management: Location and Movement

If one views supply chains as the "movement" of physical goods (see, for example, `www.ipsera.org`), then the analysis of a supply chain can focus primarily on the logistics (Razzaque and Sheng 1998) of getting goods to the ultimate consumer. The logistics research stream has focused on customer service for many years. The emphasis in this view

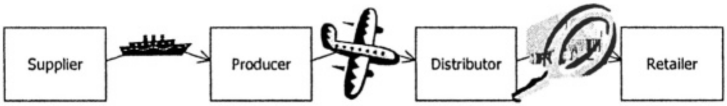


Figure 13.3. Global Logistic and Transportation View: Special Attention to the Arrows Between Partners and the Location of Distribution Points.

is on determining where actual "consumption" of the good occurs and then figuring out the optimal configuration of the network of transport arrows and transformation nodes to move raw material from the source location to the sink location of consumption. Obviously, multiple source locations and multiple (sometimes almost infinite) consumption points usually exist. Figure 13.3 provides a view of the historical emphasis in logistics on the points of transfer along the distribution channel. Kent and Flint (1997) provide a summary of the evolution of the field of logistics.

Algorithmic methods are often applied to relatively small and simple networks with limited sources, intermediate nodes, and a few consumption points. Approximation and simulation models are often applied to explore strategic decisions in more realistic and complex networks. Supply chain issues to explore can involve a very complex global environment dealing with multiple issues such as tax, duty, tariff, exchange rate, law, gaming, belief, culture, climate, politics, and so forth. As an example, if you have a product that contains a high amount of stainless steel, do you locate the transformation facilities near the mills to reduce the transport time and cost from the mill? If mills are not located near actual use of the final product, do you instead locate transformation facilities near the user nodes? The supply networks and costs will differ dramatically between these two approaches. Bowersox, Closs and Stank (2000) discuss future opportunities for logistics applications in supply chains.

3. Buyer/Supplier Dyadic Management

Buyer/supplier dyadic management research is a key carryover from years of North American research in purchasing that focused on managing first tier suppliers (`www.ISM.ws` and `www.capsresearch.org`). The emphasis is on supplier qualification, reduction in the supplier base, managing the buyer/supplier interaction, supplier development, and long-term supplier alliances (Dobler and Burt 1998). A supply chain perspective changes the focus to single sourcing, supplier pre-qualification, and specific programs by buyers to develop the capabilities of suppliers

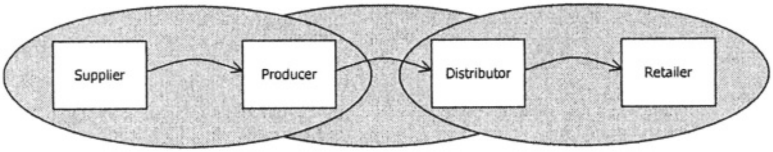


Figure 13.4. Buyer/Supplier Dyadic Management: Focus is on the Pairs within the Supply Chain.

for high quality at low cost (Forker, Ruch, and Hershauer 1999). The tendency is to view the supply chain as a series of related dyadic relationships between adjacent, overlapping pairs of supply chain partners as depicted in Figure 13.4.

The use of just-in-time (JIT) (Ohno 1988) deliveries and vendor- or supplier managed inventory (VMI or SMI) in the USA have developed in relation to this movement. JIT requires that suppliers provide components that work in the right quantity just as they are needed and require no inspection by the buyer. Time and costs are compressed for the buyer and the supplier has a regular customer at a fair price. JIT practices typically get established first on a one-on-one basis (dyadic relationship) before being implemented more systematically. SMI places the emphasis on savings and efficiency for the supplier. The supplier recruits many customers to serve using SMI. This allows the supplier to optimize logistics costs and to have full visibility to all demand. By pooling variability across buyers, stable production and minimal inventories can be maintained while minimizing transportation costs for the supplier. Benefits for the buyer stem from not having to manage the component in any manner; the component is always available at a consistently low, if not minimal, inventory cost and unit price.

Research (Choi, Ellram and Koka 2002) has now moved toward the issues surrounding triadic relationships at the first tier level (e.g., relations between one buyer and two sellers of complementary products) and extension of first tier relationships to second and third tier suppliers (Choi and Hong 2002). Extension of the conceptual and computational models beyond the dyad, even by expanding by just one more partner, has proved challenging. The introduction of the third layer highlights the potential for sub-optimization in a multi-tier transaction, and the fundamental basis for the necessary information exchange is elusive.

3.1 Communication and the Internet

Supply chains can simply be viewed as communication systems among all the stakeholders in the system. Many of the traditional problems in supply chains are the direct result of lack of a common information base and lack of timely delivery of accurate information. Information flows have long been recognized in supply chains. Towill et al. (1992) and Lee et al. (2000) described supply chains by the forward flow of materials and the backward flow of information. The central importance of information flow to supply chain management is typified by Johannson (1994): "Supply Chain Management is really an operations approach to procurement. It requires all participants of the supply chain to be properly informed. With SCM, the linkages and information flows between various members of the supply chain are critical to overall performance."

Layered atop these perspectives that stress the importance of information in the supply chain, the Internet has created a revolution in the reach and richness of information (Evans and Wurster 2000). Historically, one could have richness in communication with one other member of the supply chain or one could have reach to many with broadcasted impersonal communication. The Internet has provided the medium with minimal barrier to entry for accomplishing both richness and reach. Thus it is possible for 3,000 retail outlets, 50 retail distributors, 10 manufacturers, 100 commodity distributors, and 5 mills to all have access to exactly the same information on material and product movement and demand concurrently. The implications for managing the supply chain are enormous. The opportunities for performance improvement are huge especially in industries that have traditionally seen every member of the chain hold information very tightly, but cultural change will be required to take advantage of these technological capabilities.

3.2 Lean Supply Chains

The emphasis in lean supply chains (`www.cf.ac.uk/carbs/lom/lerc`, `www.ame.org`, and `www.lean.org`) is on extending lean manufacturing practices (Ohno 1988, Shingo 1988, Womack et al. 1990, Womack and Jones 1996, Rother and Shook 1998, Rother and Harris 2001) to first-tier suppliers and other SC participants (Jones and Womack 2002). With an emphasis on eliminating waste and managing the value stream, the main objective is working upstream with first tier suppliers and beyond to compress lead time by eliminating "flat spots." Flat spots refer to time periods with no value-adding activity.

For example, Peter Summerfield (2001), Rolls-Royce Managing Director — Transmissions & Structures, recently explained the efforts at Rolls Royce to compress total cycle time (including time to receive all materials and components, manufacture, and assemble) to produce an engine to 40 days. Rolls-Royce is compressing supplier lead time from

138 days to 40 days, manufacturing time from 54 days to 30 days, and assembly time from 42 days to 10 days. The 40, 30, and 10 day lead times are overlapped by making processes completely transparent using e-commerce, so that the total cycle time is 40 days. Using similar concepts, the time from idea acceptance to first unit has been reduced to 24 months (including a maximum of 8 months for any new manufacturing facility). Suppliers are guaranteed a viable margin and prompt payment and involvement in the next jet or engine is assumed. In addition, application of lean concepts often involves simulation analysis of flows through the supply chain. Tommelein (1998) provides an example application in the construction industry.

3.3 Process Analysis

Extending traditional process analysis and mapping across company boundaries is a suggested perspective by Hammer (2001). Damelio (1996) and Harrington (1991) provide an introduction to the tools of this research stream. The Supply-Chain Council (SCC) has developed one model for implementing process analysis across organizational boundaries. SCC is a non-profit organization comprising more than 700 member companies including manufacturers, distributors, and retailers. SCC has developed and endorsed the Supply Chain Operations Reference-model (SCOR) as the cross-industry standard for supply chain management. The SCOR model includes four process levels. Information on their web site on July 17, 2003 indicates that the "Top Level" contains the five process types of "Plan, Source, Make, Deliver, and Return." The "Configuration Level" then specifies core "Process Categories." The "Process Element Level" shows how processes are decomposed into elements with information inputs and outputs. At this level, performance metrics and best practices are established. The "Implementation Level" shows how process elements are further decomposed to define practices to achieve competitive advantages in the supply chain. More information on the Supply Chain Operations Reference (SCOR) Model can be obtained from the Supply Chain Council (`www.supply-chain.org`). Organizations in the SCC are seeking to improve supply chain performance in terms of speed, increased reliability, lower operating cost, and lower inventories; all of these objectives are consistent with the lean objectives.

In addition to emphasizing reduction of lead time and inventory levels, the lean perspective reaches into many aspects of the production process in the supply chain. The lean viewpoint is about the aggressive eradication of waste in the supply chain, where waste is cast very broadly as a failure to meet customer requirements. With this definition as a back-drop, the lean approach can also lead to reconfiguration of the product design and assembly process itself, for instance to make the good more amenable to postponement of customization. This holistic approach to managing the supply chain and the production process it represents can be extremely effective.

3.4 Power and Relationships

Working from basic economic theory, Andrew Cox evaluates supply chains through the lens of market maturity and power (see, for example, www.business.bham.ac.uk/business/page539.htm). This supply chain perspective changes the focus from leveraging existing suppliers, to building trust and collaboration, to evaluating critical assets and who has power over supply chain resources, and finally to reconfiguring supply chain power through strategic innovation and realignment (Cox 1999). Under this view, the primary analytical component is at the strategic level for a specific market or industry (e.g., www.m4i.org.uk/). Existing boundaries of firms may become relatively porous and flexible before being subjected to strategic realignment resulting from disintermediation and deconstruction. Cox appears to be relatively uninterested in the logistics or production processes, or in the inventory levels anywhere in the supply chain. Rather, he espouses a philosophy that strong relationships up and down the supply chain will lead to effective and powerful supply chain performance. His disinterest in the particular technologies used within the supply chain grows out of his belief that reconfiguration of the supply chain, and not introduction of new technologies or processes, will ultimately lead to the most dramatic and sustainable competitive advantages for supply chains in their entirety.

Today's organizations are changing rapidly in order to remain competitive. They are, accordingly, reshaping their own supply chains. Davis and Spekman (2004) talk of gaining competitive advantage through "the extended enterprise" that requires people to work across organizational boundaries. Achieving the connectivity, community, and collaboration that Davis and Spekman espouse will require new forms of sharing power and control of the supply chain. Change often happens out of necessity, when an organization faces a crisis. It is understandable that successful businesses prefer a status quo to a change. Change is always perceived as risky and maintaining a status quo may not be perceived as such. Nevertheless, there always are some companies within any industry that have to face change, be it out of necessity or opportunity, and their industry as a whole will therefore be subject to change.

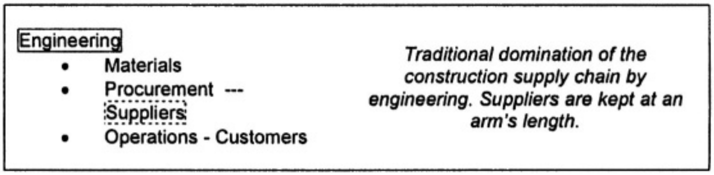

Engineering
• Materials
• Procurement --- Suppliers
• Operations - Customers
Traditional domination of the construction supply chain by engineering. Suppliers are kept at an arm's length.

Figure 13.5. London Underground's Supply Chain pre 1988 (Adapted from Figure 7.1 in Lamming and Cox 1995).

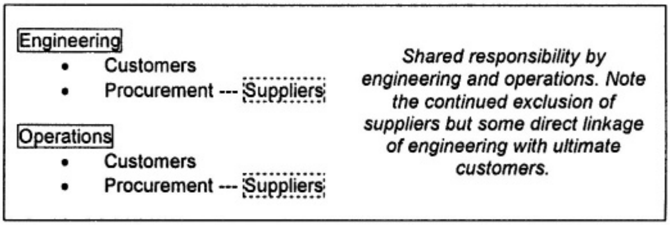

Engineering
• Customers
• Procurement --- Suppliers
Operations
• Customers
• Procurement --- Suppliers
Shared responsibility by engineering and operations. Note the continued exclusion of suppliers but some direct linkage of engineering with ultimate customers.

Figure 13.6. London Underground's Supply Chain post Kings Cross (Adapted from Figure 7.2 in Lamming and Cox 1995).

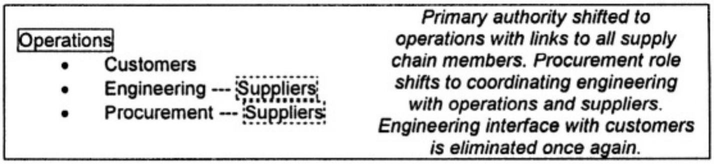
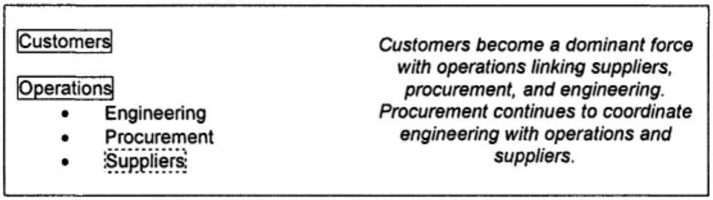

Operations
• Customers
• Engineering --- Suppliers
• Procurement --- Suppliers
Primary authority shifted to operations with links to all supply chain members. Procurement role shifts to coordinating engineering with operations and suppliers. Engineering interface with customers is eliminated once again.

Figure 13.7. London Underground's Supply Chain Post Company Plan (Adapted from Figure 7.3 in Lamming and Cox 1995).

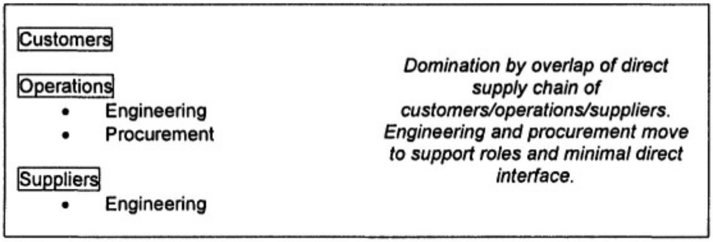
To illustrate the dramatic changes organizations may encounter, consider the case of the London Underground. Bouverie-Brine and Macbeth (Chapter 7 in Lamming and Cox 1995) described the change in thinking and the resulting realignment within their organization as well as the realignment of their organization's supply chain over a 10-year period. Figures 13.5 through 13.9 depict these changes cogently. Their organization continues to undergo change in order to better meet customer demand. Comments added to the figures reflect an interpretation of the findings. Supply chain relationships are seldom constant as responsibilities and power in the chain instigate reconfiguration.

Organizational change takes place over an extended period of time; the adoption of supply chain management practices also is done in phases.



Customers become a dominant force with operations linking suppliers, procurement, and engineering. Procurement continues to coordinate engineering with operations and suppliers.

Figure 13.8. London Underground's Supply Chain Post Introduction of Supplier Managers (Adapted from Figure 7.5 in Lamming and Cox 1995).



Domination by overlap of direct supply chain of customers/operations/suppliers. Engineering and procurement move to support roles and minimal direct interface.

Figure 13.9. London Underground's Supply Chain of the Future (Adapted from Figure 7.8 in Lamming and Cox 1995).

3.5 Mass Customization

Another way to view developments in supply chains is through the lens of mass customization(`www.mass-customisation.org`).The basic concept in mass customization is to produce and deliver products and services that are unique and customer-specific with the same speed and efficiency as mass-produced standard products. Marketing and advertising as they are known today could die completely if this model dominates since there would be no standard product to market, but no one really expects this to happen completely. Nonetheless, historically supply chains in many industries were relatively stable. Lampel and Mintzberg (1996) provide an overview of the continuum from pure standardization to pure customization. They argue that historically stable markets and historically unique markets are moving toward a middle ground. The construction industry provides an important exception, wherein customized products — the facilities — are the rule rather than the exception (Tommelein et al. 2003). The pace of technological and product innovation has rendered many supply chains more ephemeral. Rapid reconfiguration and postponement of customization are often key enablers to long-term competitiveness of 'on-demand' supply chains.

By focusing on ultimate consumer or customer specifications to meet need, by fulfilling orders using a pull structure, and by having flexible and agile operations and logistics, the emphasis is on delivering quality at low cost in a rapid response environment. These practices warrant the use of the term ‘demand chain’ rather than ‘supply chain.’ They require full transparency upstream from the ultimate customer through the chain and short response times. Often, emphasis is placed on postponement of customer-specific modifications as far downstream as possible so that standardization can be accomplished upstream. Concurrent engineering is a must in this environment. Even the customer is involved in new product idea generation.

3.6 Relationships and Alliances

Under the relationships and alliances view, the emphasis is on gathering the right group of players together to strike out on new approaches to a demand or supply chain. The emphasis is on strategic partnering and working relationships (Hammer 2001) rather than obsessive concern for merger, acquisition, or other traditional forms of reconfiguration. Internal alliances may be just as important as external ones in many large corporations.

An alliance is a long-term relationship agreement between parties to continuously improve work processes and the reliability of performance. It is characterized by joint problem solving and process improvement and by trust, respect, cooperation, and mutual benefits. Alliance agreements lock in favorable or predictable pricing and delivery terms. They usually include agreements regarding the exchange of information among the alliance partners. Typically, alliance agreements might be set up with a small percentage of a company’s suppliers.

Strategic products include long-lead, complex, big-ticket items or key components, whereas strategic services usually represent significant intellectual capital. Figure 13.10 shows these in the upper-right quadrant of the sourcing square. Customers will be inclined to develop strong, long-term agreements with these suppliers.

For products and services that are critical to the company’s business, but that have low cost, a company may set up company-wide purchasing agreements. Figure 13.10 shows these in the upper-left quadrant of the sourcing square. Pricing agreements typically specify types of products, minimum and maximum volumes to be purchased, delivery, and price. They could be company, project, or program specific. Figure 13.10 shows these in the lower-right quadrant of the sourcing square. Spot market purchases are made on an as-needed basis. These purchases

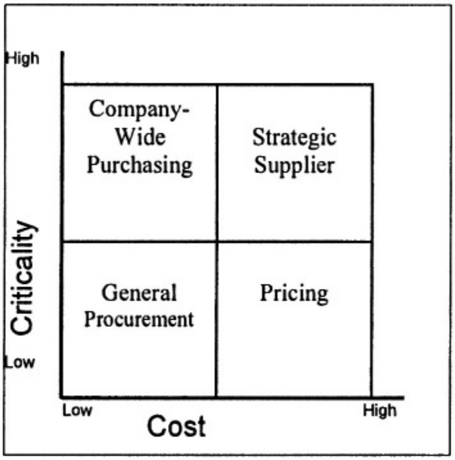


Figure 13.10. Sourcing Agreement Square (Adapted from Figure 10.2 in Banfield 1999).

assume that the product or service is widely and readily available. Figure 13.10 shows these in the lower-left quadrant of the sourcing square. Different agreements lead to different functions being filled by supply chain partners.

According to Segil (2001 p. 23) "the trend is toward large numbers of alliances that are nonexclusive rather than small numbers of alliances that are exclusive." Virtual networks and supply chains can be formed and reformed at a rapid rate to find new markets, revolutionize old markets, or become a new supply chain alternative to a traditional market. Informal cooperation and consortium arrangements may be more prevalent. For example, an intermediate manufacturer and a software company may join forces to develop collaboration tools for a particular supply chain. This may involve sharing employees whether or not equity positions are established.

An intriguing approach to creating new relationships are the private finance initiatives concerning highway construction. By changing to purchasing a flow of services from capital assets rather than the capital asset itself, alliances of suppliers have formed into companies to design, build, finance, and operate (DBFO) highways. "Where previously suppliers' attention was paid to claims positioning and the potential for leveraging profit through the identification of liabilities, under the DBFO problem-solving attitudes overrode those of problem avoidance" (Hall et al. 2000 p. 227).

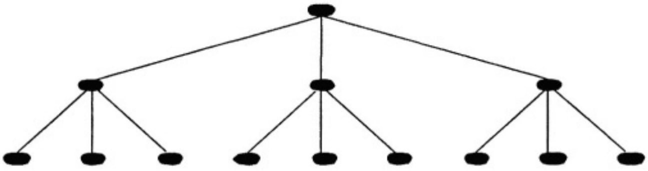


Figure 13.11. Hierarchy of Buyer Levels.

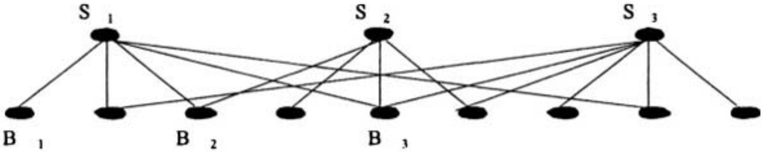


Figure 13.12. Market relationships between buyers and sellers.

3.7 Information Access, Market Structures, and Control

Looking at market structures for supply chains from roots in economic and organization theories (for example, see Vrijhoef et al. 2003), one can envision different substructures for physical transactions and information transactions as (1) hierarchical structures, (2) centralized markets, (3) decentralized markets, and (4) hybrid structures (Malone 1987). On one hand, distribution structures (business to business) may be hierarchical: firms use only one supplier for a particular good (Halal 1994). This choice is often based on a customer's desire to minimize search costs. Figure 13.11 depicts a typical hierarchy with many buyers for each supplier.

On the other hand, customer strategies may use both hierarchies and markets. For example, some customers may buy a particular part at one supplier and never consider using multiple sources or an alternate supplier. The same customer may choose to shop around and consider many outlets (conventional- and electronic storefronts) in purchasing another particular part. This desire to use the market mechanism requires a significant expenditure of time and effort. However, this strategy gives a customer access to a number of suppliers thus resulting in a market type relationship between a group of buyers and sellers. Thus, a one-to-many relationship between the two groups results for some buyers as part of a many to many market structure as depicted in Figure 13.12.

Figure 13.12 illustrates different types of customers: B1 is the loyal buyer who sticks to one supplier; B2 considers two; B3 uses three.

Within this example, customers can be connected to the distributors in a hierarchy-like fashion such as B1 (one upward connection). However, other customers make contacts with numerous suppliers (such as B2 or B3) thus changing the basis of connectivity and information use and roles by buyers and sellers. These buyers choose to operate in the market environment to minimize a specific objective such as price or capacity availability. Another organizational structure that favors electronic connectivity and access is an electronic market, which is a form of a centralized market. This type of structure is characterized by a set of intermediaries between buyers and sellers and high degree of connectivity between the three layers (Talalayevsky and Hershauer 1997). Not all buyers are equally motivated to switch from the conventional structures into electronic markets because they possess different degrees of information about the purchasing decisions. In other words, given the buyers current information about products and access to potential sellers, it may not be worthwhile for many to expand the effort of search (cost) to explore new electronic structures.

Many supply chain changes are the result of greater access to complete information and use of information brokers for control and efficiency of the market. Hybrid structures are enabled by information technology and can be understood through the lens of coordination theory (Lewis and Talalayevsky 1997). Hybrid structures tend to use third-party information brokers, exchanges, auctions, and so forth to move markets to commodity relationships where possible and to very specialized hierarchical markets with dynamic alliances (Choi, Dooley and Rungtusanatham 2001) for unique needs.

3.8 Environmental Life Cycle

Environment is a global issue. Although many concerns about the environmental impact of manufacturing have focused on local and regional issues, the larger issues of resource and energy sustainability, irreversibility of some chemical changes, global climate and ecological system changes, and differential responsibilities and economic realities create the need for a truly global perspective regarding supply chains and the environment. Supply chain issues and decisions become very complex if long-term impact and cost implications are considered at the global level.

Based on concepts and models in Caporello (1995), Daniel, et al. (1997), Hershauer (1994), and Wu and Dunn (1995), a conceptual model of an expanded life cycle analysis for operations managers and the supply loop might incorporate environmental issues and management practices as shown in Table 13.1. References provided in this section contain definitions and discussions of the environmental terms such as clean production, takeback, reverse logistics, and open recycling. Although the key environmental issue and the corresponding key management response are based on an extensive review of practice and literature, they are certainly not the only possible items to list. They are provided to underline the changes and complexity caused by considering environmental issues in an expanded supply loop. The term loop is used to indicate that the materials upstream in the supply chain eventually become the materials upstream in another chain.

Product Stage	Key Environmental Issue	Key Management Response
1. Product Definition	Consumer preferences	Creativity/innovation
2. Product Development	Technology Scanning	Design For Environment
3. Process Design	Zero toxic waste goal	Clean production
4. Raw material	Unused material	Reuse of components
5. Transformation	Wastes	Source reduction
6. Assembly	Scrap	Yields & resource use
7. Product Delivery	Packaging	Identify for tracing
8. Use	Returns	"Ownership"
9. Retire	Reverse logistics	Managing uncertainty
10. Return	Product takeback	Reuse of products
11. Second user	Returns	Global redistribution
12. Disassembly	Wastes and scrap	Design For Disassembly
13. Reuse	Replaced parts	Remanufacturing processes
14. Assembly	Scrap	Tracing quality
15. Product Delivery	Packaging	Distribution ethics
16. Second Use	Returns	Alternate Use Innovations
17. Reduce	Wastes	Open recycling
18. Recycle r.m.	Excess r.m.	Storage

Table 13.1. Supply Loop Stages.

Future supply chain managers will need to focus more on how to make tactical operating decisions in response to major strategies such as full product takeback, design and manufacture for disassembly, design and manufacture for zero toxic waste disposal, manufacturing new products with mostly reused and remanufactured components, manufacturing without depleting resource reserves, and designing and manufacturing to minimize energy usage over a product's entire life cycle of creation, use, reuse, downcycling, and recycling. Life cycle analysis (Caporello, 1995 and Miettinen and Hamalainen, 1997) will become a common ingredient to product idea generation, product design, product development, first unit production, and full production and distribu-

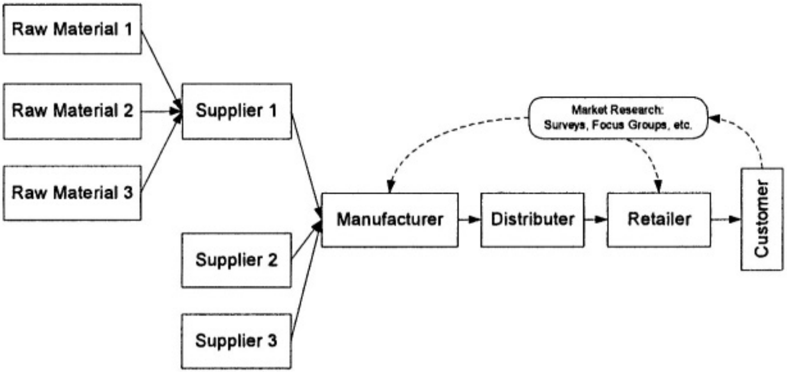


Figure 13.13. Typical Retail Supply Chain (Adapted from Simchi-Levi, et al. 2000).

tion. Operational research models (Maloni and Benton, 1997) that assist in analyzing tradeoffs and interactions amongst the 18 stages, environmental issues, and management practices identified above are needed if systemic effects are to be correctly and comprehensively known.

4. Industry Influences

The supply chain literature commonly presents supply chains that have a linear organization. In fairness, such a depiction simplifies many of the concepts of supply chain management, and for that reason the preceding figures in this chapter have been based on such simplification. These diagrams tend to highlight the length and coordination aspects of supply chain management, but tend to de-emphasize the selection of supply chain participants at each tier, the expanding nature of successive tiers, and the interactions between the customer and the numerous other supply chain participants. Figure 13.13 shows a modification to the traditional conceptual model of the supply chain that highlights some of these factors more directly. Supply chains in many industries may have, at their fringes, subset supply chains that resemble Figure 13.13, but the overall structure of the supply chain and the degree and manner of customer interaction with the supply chain may be radically different. A few examples are presented by way of illustration in Figures 13.14, 13.15, and 13.16. These figures are not intended to be all-encompassing; there are procurement strategies or delivery methods in each industry that might have different structures. The intention here is to be illustrative rather than exhaustive.

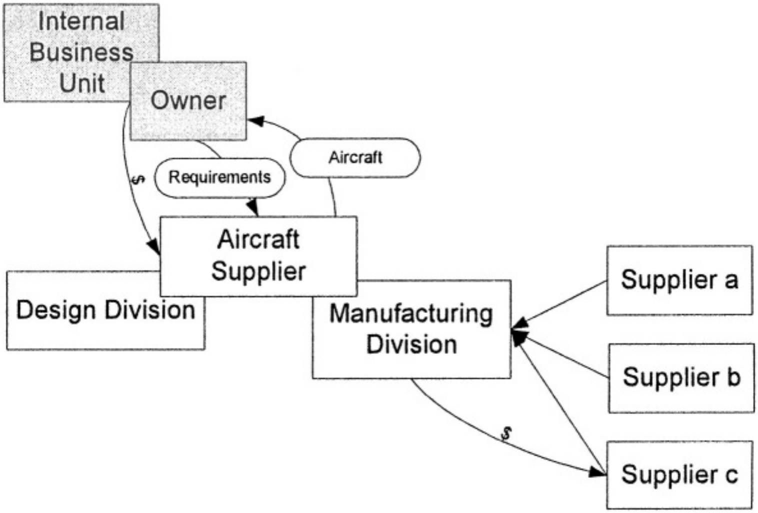


Figure 13.14. Typical Aircraft Supply Chain.

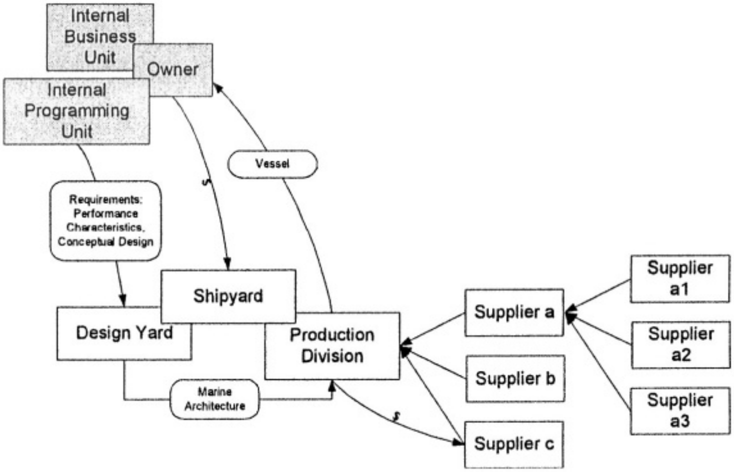


Figure 13.15. Typical Marine Supply Chain.

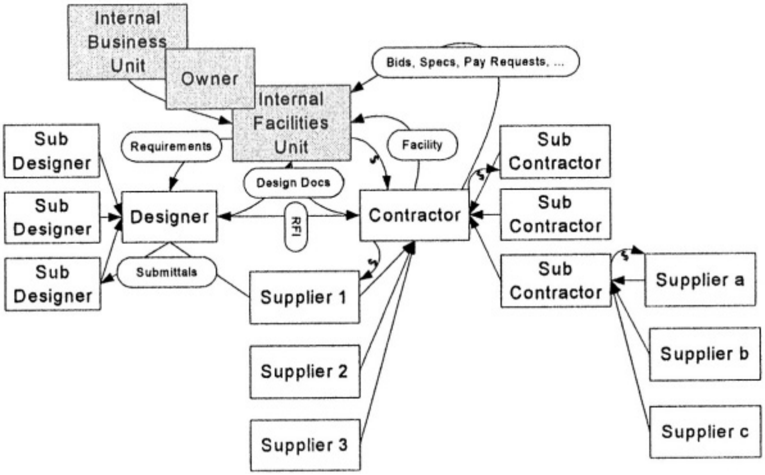


Figure 13.16. Typical Design-Bid-Build Construction Supply Chain.

The structure of the supply chain and the degree and manner of customer interaction with the supply chain influence the viewpoints typically taken by researchers within a particular industry. Put another way, it seems to us that the development of a predominant view for researchers in a particular industry may be as influenced by "nature" (by which we mean the inherent appropriateness of a particular view to a particular set of conditions) as it is by "nurture" (the academic tradition in which the researchers developed). Or, it may be that researchers who come from a perspective especially relevant to an industry tend to be attracted to that industry. Whether it is the egg or the chicken that comes first, it is noticeable that some views tend to achieve more importance in some industries. Where the customer interaction is relatively indirect (Figure 13.13), emphasis on the length, coordination, and production levels tend to become relatively more influential (the Inventory view may dominate, for example). By contrast, where the customer interaction occurs at multiple tiers and is pervasive in the process (Figure 13.16) perspectives emphasizing the importance of the relationship (the power and relationship research view, for example) tend to become relatively more important.

5. Summary of Research Streams

This chapter has included brief descriptions of several different research approaches to supply chain management. While there are differences among these various approaches, there are many overlaps and similarities as well.

Table 13.2 is provided to depict the differences, overlaps, and similarities in approach in more detail. Here, the primary and secondary concerns relating to each point of view are indicated. The lack of indication of a primary or a secondary concern does not necessarily imply that this characteristic of the supply chain is ignored in that view, only that it is not one of the chief concerns.

View	Interface Between Parties	Inventory Levels	Process Improvement	Product Development	Communication Systems	Shipping and Transportation	Equity Arrangements
1. Inventories		●			○		
2. Global Logistics and Management						●	
3. Buyer-Supplier Dyadic Management	●	○			○	○	
4. Communication and the Internet					●		
5. Lean Supply Chains		○	●	○	○	○	
6. Process Analysis	○	○	●		○	○	
7. Power and Relationships	●						○
8. Mass Customization			○	●			
9. Relationships and Alliances	●				○		○
10. Information Access, Market Structures, and Control	○				●		○
11. Environmental Life Cycle	●	○	○	●	○	○	○

Table 13.2. Summary of Areas of Concern in Each View. Primary concerns indicated with solid circle (•), secondary concerns with open circle (○).

6. Concluding Thoughts

These eleven views provide a window into the vast literature and array of views about supply chains. Cox (1997) provides an evaluation that is helpful in understanding recent and current activity in supply chains. Basically a simple statement of level of maturity in supply chain management consideration, Cox's framework provides a clear picture that practice and literature reflects mostly early stage consideration by managers over the last decade. At the lowest level, key operational questions include:

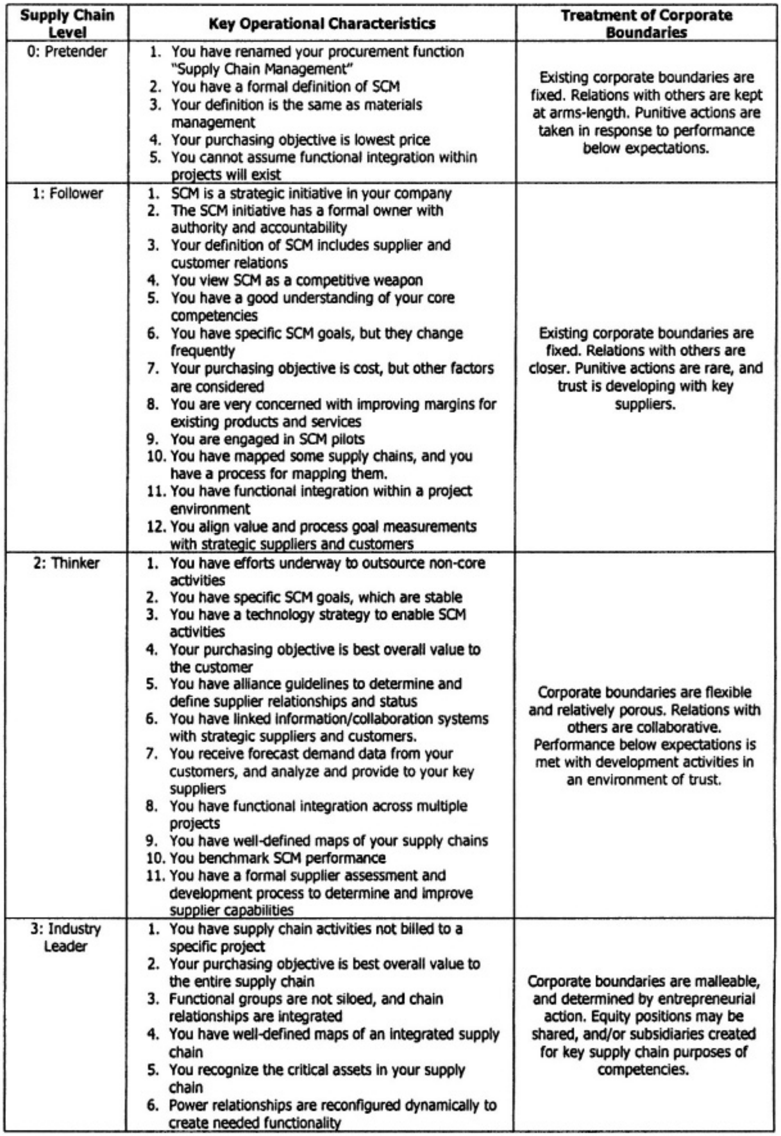
- "How can existing suppliers be leveraged?"
- "How can existing products be aligned with supply chains appropriately?" (Cox, 1997, p. 299)

These questions are typical of those commonly asked in management practice in industry over the last decade. Companies considering supply chain initiatives at present generally express interest in the second level questions presented by Cox:

- "How can value be delivered to the customer more effectively?"
- "How can trust and collaboration be built?"
- "How can virtual organizations be maintained?" (Cox, 1997, p. 299)

As a final commentary on the literature and its application by organizations, Table 13.3 highlights the phases of maturity in the adoption of supply chain concepts. Many firms in industry fit into the pretender and follower categories. Some are in or close to the thinker category. There are probably no firms in the leader category at this time.

Acknowledgements: Writing of this overview of schools of thought in supply chain management was funded in part by a grant for Project Team 172 "Improving Construction Supply Chain Performance" from the Construction Industry Institute whose support is gratefully acknowledged. Any opinions, findings, conclusions, or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of the Construction Industry Institute.



Supply Chain Level	Key Operational Characteristics	Treatment of Corporate Boundaries
0: Pretender	1. You have renamed your procurement function "Supply Chain Management" 2. You have a formal definition of SCM 3. Your definition is the same as materials management 4. Your purchasing objective is lowest price 5. You cannot assume functional integration within projects will exist	Existing corporate boundaries are fixed. Relations with others are kept at arms-length. Punitive actions are taken in response to performance below expectations.
1: Follower	1. SCM is a strategic initiative in your company 2. The SCM initiative has a formal owner with authority and accountability 3. Your definition of SCM includes supplier and customer relations 4. You view SCM as a competitive weapon 5. You have a good understanding of your core competencies 6. You have specific SCM goals, but they change frequently 7. Your purchasing objective is cost, but other factors are considered 8. You are very concerned with improving margins for existing products and services 9. You are engaged in SCM pilots 10. You have mapped some supply chains, and you have a process for mapping them. 11. You have functional integration within a project environment 12. You align value and process goal measurements with strategic suppliers and customers	Existing corporate boundaries are fixed. Relations with others are closer. Punitive actions are rare, and trust is developing with key suppliers.
2: Thinker	1. You have efforts underway to outsource non-core activities 2. You have specific SCM goals, which are stable 3. You have a technology strategy to enable SCM activities 4. Your purchasing objective is best overall value to the customer 5. You have alliance guidelines to determine and define supplier relationships and status 6. You have linked information/collaboration systems with strategic suppliers and customers. 7. You receive forecast demand data from your customers, and analyze and provide to your key suppliers 8. You have functional integration across multiple projects 9. You have well-defined maps of your supply chains 10. You benchmark SCM performance 11. You have a formal supplier assessment and development process to determine and improve supplier capabilities	Corporate boundaries are flexible and relatively porous. Relations with others are collaborative. Performance below expectations is met with development activities in an environment of trust.
3: Industry Leader	1. You have supply chain activities not billed to a specific project 2. Your purchasing objective is best overall value to the entire supply chain 3. Functional groups are not siloed, and chain relationships are integrated 4. You have well-defined maps of an integrated supply chain 5. You recognize the critical assets in your supply chain 6. Power relationships are reconfigured dynamically to create needed functionality	Corporate boundaries are malleable, and determined by entrepreneurial action. Equity positions may be shared, and/or subsidiaries created for key supply chain purposes of competencies.

Table 13.3. Categories of Supply Chain Leadership After Cox, A. (1997), Business Success: A Way of Thinking About Strategy, Critical Supply Chain Assets and Operational Best Practice, Earlsgate Press: Bath, U.K.

References

Banfield, E. (1999). Harnessing Value in the Supply Chain: Strategic Sourcing in Action. John Wiley & Sons, Inc., New York .

Bouverie-Brine, C. and Macbeth, D.K. (1995). Managing Supply Chains: A Collaborative Project between London Underground and the Supply Chain Management Group. Chapter 7 pp. 115-127 in Lamming and Cox (1995).

Bowersox, D.J., Closs, D.J. and Stank, T.P. (2000). Ten Mega-Trends that Will Revolutionize Supply Chain Logistics. Journal of Business Logistics 21 (2), 1-16.

Caporello, T. J. (1995). A design for the environment advisor for product and process design selection. Unpublished doctoral dissertation, Arizona State University.

Champy, J. (2002). X-Engineering the Corporation. Warner Books, Inc., New York.

Choi, T.Y., Dooley, K.J., and Rungtusanatham, M. (2001). Supply networks and complex adaptive systems: control versus emergence. Journal of Operations Management 19 (3), 351-366.

Choi, T.Y., Hong, Y. (2002). Unveiling the structure of supply networks: case studies at Honda, Acura, and DaimlerChrysler. Journal of Operations Management 20 (5), 469-493.

Choi, T.Y., Wu, Z., Ellram, L., and Koka, B. R. (2002). Supplier-Supplier Relationships and Their Implications for Buyer-Supplier Relationships. IEEE Transactions on Engineering Management, 49 (2), 119-130.

Churchman, C.W. (1961). Prediction and Optimal Decision. Prentice-Hall, Englewood Cliffs, N.J.

Cox, A. (1997). Business Success: A Way of Thinking About Strategy, Critical Supply Chain Assets and Operational Best Practice, Earlsgate Press, Bath, U.K.

Cox, A. (1999). Power, Value and Supply Chain Management. Supply Chain Management Journal: An International Journal, 4 (4), 167-175.

Cox, A. and Townsend, M. (1998). Strategic Procurement in Construction. Thomas Telford, London.

Damelio, R. (1996). The Basics of Process Mapping. Productivity, Inc., Portland, OR.

Daniel, S.E., Diakoulaki, D.C., Pappis, C.P. (1997). Operations Research and Environmental Planning. European Journal for Operational Research, 102 (2), 248-263.

Davis, E.W. and Spekman, R.E. (2004). The Extended Enterprise. FT Prentice Hall, Upper Saddle River, NJ.

Dobler, D.W. and Burt, D.N. (1998). Purchasing and Supply Management: Text and Cases, Sixth Edition, McGraw-Hill, Inc., Boston.

Evans, P. and Wurster, T.S. (2000). Blown to Bits. Harvard Business School Press, Boston.

Forker, L. Ruch. W. and Hershauer, J. (1999). Examining Supplier Efforts from Both Sides. The Journal of Supply Chain Management, 35 (3), 40-50.

Forrester, J.W. (1961). Industrial Dynamics. MIT Press, Cambridge (currently available from Pegasus Communications, Waltham, MA).

Forrester, J.W. (1969). Urban Dynamics. Pegasus Communications, Waltham, MA.

Forrester, J.W. (1971). World Dynamics. Pegasus Communications, Waltham, MA.

Goldratt, E.M. and Cox, J. (1986). The Goal. North River Press, Croton-on-Hudson, NY.

alal, W. E. (1994). From the Hierarchy to Enterprise: Internal Markets Are the New Foundation of Management. Academy of Management Executive, 8 (4), 69-83.

Hall, R., Holt, R. and Graves, A. (2000). Private Finance, Public Roads: Configuring the Supply Chain in PFI Highway Construction. European Journal of Purchasing & Supply Management, 6, 227.

Hammer, M. (2001). The Agenda: What Every Business Must Do to Dominate the Decade. Crown Business, New York.

Harrington, H.J. (1991). Business Process Improvement. McGraw-Hill, Inc., New York.

Hershauer, J.C. (1994). Elements of an Environmental Compliance Infrastructure (report on a talk by Ralph Ponce de Leon, Motorola, Inc.). Operations Management Review 10 (3), 51-54.

Johannson, L. (1994). How Can a TQEM Approach Add Value to Your Supply Chain. Environmental Quality Management. Summer, 3 (4), 521-530.

Jones, D. and Womack, J. (2002). Seeing the Whole - Mapping the Extended Value Stream. March, The Lean Enterprise Institute, Brookline, MA.

Kent, J.L. and Flint, D.J. (1997). Perspectives on the Evolution of Logistics Thought. Journal of Business Logistics 18 (2), 15-29.

Lamming, R. and Cox, A. (eds.) (1995). Strategic Procurement Management in the 1990s. Concepts and Cases. Earlsgate Press, Great Britain.

Lampel, J. and Mintzberg, H. (1996). Customizing Customization. Sloan Management Review, Fall 1996, 21-30.

Langdon, D.S., Richelsoph, J., and Martin, T. (1999). Purchasing Performance Benchmarks for the U.S. Engineering/Construction Industry. Tempe, Arizona, Center for Advanced Purchasing Studies, Arizona State University.

Laufer, A. (1997). Simultaneous Management. American Management Association, New York.

Lee, H. L., So, K. C., and Tang, C. S. (2000). The Value of Information Sharing in a Two-Level Supply Chain, Management Science, 46 (5), 626 - 643

Lewis, I. and Talalayevsky, A. (1997). Logistics and Information Technology: A Coordination Perspective. Journal of Business Logistics, 18 (1), 141-157.

Luhtala, M., Kilpinen, E., and Anttila, P. (1994). ALOGI: Managing Make-to-Order Supply Chains. Helsinki University of Technology, Espoo, Finland.

Malone, T. W. (1987). Modeling Coordination in Organizations and Markets, Management Science, 33 (10), 1317-1331.

Maloni, M.J., Benton, W.C. (1997). Supply Chain Partnerships: Opportunities for Operations Research. European Journal for Operational Research 101 (3) 419-429.

Melnyk, S.A. and Swink, M. (2002). Value-Driven Operations Management: An Integrated Modular Approach. McGraw-Hill/Irwin, New York.

Miettinen, P., Hamalainen, R.P. (1997). How to Benefit from Decision Analysis in Environmental Life Cycle Assessment (LCA). European Journal for Operational Research 102 (2), 279-294.

Miller, D.W. and Starr, M.K. (1960). Executive Decisions and Operations Research. Prentice-Hall, Englewood Cliffs, N.J.

Ohno, T. (1988). Just-in-time for Today and Tomorrow. Taiichi Ohno with Setsuo Mito; translated by Joseph P. Schmelzeis, Productivity Press, Cambridge, MA.

Razzaque, M.A. and Sheng, C.C. (1998). Outsourcing of Logistics Functions: A Literature Survey. International Journal of Physical Distribution & Logistics Management, 28 (2), 89-107.

Rother, M. and Harris, R. (2001). Creating Continuous Flow - An Action Guide for Managers, Engineering, and Production Associates. V. 1.0, June, The Lean Enterprise Institute, Brookline, MA.

Rother, M. and Shook, J. (1998). Learning to See: Value Stream Mapping to Add Value and Eliminate Muda. V. 1.1, October, The Lean Enterprise Institute, Brookline, MA.

Segil, L. (2001). Fast Alliances. John Wiley & Sons, New York.

Shingo, S. (1988). Non-stock Production: the Shingo System for Continuous Improvement. Productivity Press, Cambridge, MA.

Simchi-Levi D., Kaminsky P., and Simchi-Levi E. (2000). Designing and Managing the Supply Chain: Concepts, Strategies, and Case Studies. McGraw-Hill/Irwin, New York.

Sloan, Jr., A.P. (1963). My Years with General Motors. Doubleday, Garden City, New York.

Summerfield, P. (2001). Plenary Speech by Rolls-Royce Managing Director - Transmissions & Structures, What Really Matters in Operations Management, European Operations Management Association, Bath, UK.

Talalayevsky, A. and Hershauer, J.C. (1997). Organizational Coordination Cost Evaluation of Network Configurations. Journal of Organizational Computing and Electronic Commerce, 7 (2 & 3), 185-199.

Tommelein, I.D. (1998). Pull-driven Scheduling for Pipe-spool Installation: Simulation of Lean Construction Technique. ASCE, Journal of Construction Engineering and Management 124 (4), 79-88.

Tommelein, I.D., Riley, D., and Howell, G.A. (1999). Parade Game: Impact of Work Flow Variability on Trade Performance. ASCE, J. of Constr. Engrg. and Mgmt., 125 (5), 304-310.

Tommelein, I.D., Walsh, K.D., and Hershauer, J.C. (2003). Capital Projects Supply Chain Management. RR172-11, Construction Industry Institute, Austin, Texas.

Towill, D. R., Naim, M. M., Wikner, J. (1992). Industrial Dynamics Simulation Models in the Design of Supply Chains. International Journal of Physical Distribution & Logistics Management, 22 (5), 3-13.

Vrijhoef, R. and Koskela, L. (2000). The Four Roles of Supply Chain Management in Construction. European J. of Purchasing & Supply Chain Mgmt., 6, 169-178.

Vrijhoef, R., Koskela, L. and Voordijk (2003). Understanding Construction Supply Chains: A Multiple Theoretical Approach to Inter-Organizational Relationships in Construction. Proc. 11th Annual Conference of the International Group for Lean Construction, J. C. Martinez and C. T. Formoso (editors) Blacksburg, VA, 22-24 July 2003, 280-292.

Womack, J.P. and Jones, D.T. (1996). Lean Thinking: Banish Waste and Create Wealth in your Corporation. Simon and Schuster, New York.

Womack, J.P., Jones, D.T., and Roos, D. (1990). The Machine that Changed the World. Harper Collins, New York.

Wu, H.-J. and Dunn, S.C. (1995). Environmentally Responsible Logistics Systems. International Journal of Physical Distribution & Logistics, 25 (2), 20-38.

Chapter 14

PROFIT: DECISION TECHNOLOGY FOR SUPPLY CHAIN MANAGEMENT AT IBM MICROELECTRONICS DIVISION

Ken Fordyce, Gerald (Gary) Sullivan
IBM Microelectronics Division
1000 River Road
Essex Junction, Vermont
fordyce@us.ibm.com

Abstract The primary purpose of supply chain management applications is helping an organization respond to events in a synchronized and timely fashion. During the 1990s most research and development work focused on improving the level of centralized (and at times optimal) control. This was and is a huge task and some remarkable successes have been achieved. As with any science, the accomplishment of one goal not only brings a sense of pride, but a huge dose of reality in what is left to achieve. In supply chain management (SCM), achieving reasonable levels of strong central control has dramatically increased organizational performance, but clearly identified gaps in timely synchronized response that can only currently be handled with ad hoc manual intervention that operates without global awareness. Achieving the next leap in SCM support requires harnessing collaborative tools. This chapter explores these issues through a study of recent SCM efforts in support of IBM's Technology Group.

1. Introduction

Every organization is faced with the challenge of responding to events in a synchronized and timely fashion. The response may range from a decision not to change the current state of execution to a radical overhaul. For example, a soccer team must constantly adjust as the ball and players change position. Each player must change in co-ordination with the other players and the team game plan. Typically, the competitor with best team intelligence will win the match.

In the supply chain management arena most of the 1990s was focused on improving the level of centralized (and at times optimal) control. This was and is a huge task and some remarkable successes have been achieved. This work required creating a unified (data) image of an organization and building and implementing decision technology models that could build a competent unified plan in a reasonable period of time.

As with any science, the accomplishment of one goal, not only brings a sense of pride, but a huge dose of reality in what is left to achieve. In supply chain management (SCM), achieving reasonable levels of strong central control has dramatically increased organizational performance, but clearly identified gaps in timely synchronized response that can only currently be handled with ad hoc manual intervention that operates without global awareness. A simple example is a change in an order once the plan has been established. A second example is a component can be finished earlier then planned cycle times to meet an order, but no connected between the two activities exists. The flip side of a GAP is an opportunity window.

A major opportunity window now exists for decision technology within SCM to move from tools that operate once a time period (a week) to an on demand or real-time frame. A variety of terms have begun to show up: intelligent agents (IA), supply chain event management (SCEM), and SaR (sense and respond) to name three. To achieve success here, we must intertwine collaboration and centralization. NEITHER by ITSELF is up to the challenge. It is similar to the challenges faced by factory scheduling community in the 1980s and early 1990s (Zweben and Fox 1994).

Effective centralization refers to the ability to take into consideration all aspects of the decision situation simultaneously and generate one optimal or at least very good solution. To be effective a centralized solution requires a synchronized current view of the entire decision landscape, the ability to deal with complex trade-offs, and fast performance. Gaps exist: generated by time lags, summarization, performance, triggers, and formulation. By trigger we refer to the event that wakes up and runs the central solution (typically a date). Once a week, once a day, once every 3 days, the central solver executes. The decision to execute is made without any knowledge or monitoring of events since the last execution. Formulation gap refers to the inability to formulate key decision questions in a manner that lends itself to a central solution as opposed to sequence of negotiation or collaborations.

Collaboration refers to an iterative process that focuses on finding a satisfactory solution. The next search process step depends on prior steps and may involve back tracking. Often the step involves negotiating a temporary change in the rules governing the game and typically contingency occurs to handle uncertainty. Typically current "gap limitations" preclude a centralized solution.

The goal of the chapter is to provide a concise case history of the ongoing evolution of a major SCM effort in support of IBM's Technology Group to help characterize the scope and scale of such application, identify potential opportunities for improvement and set them within a logical evolutionary pattern, and identify research opportunities to develop new decision support capabilities.

2. Brief History and Current Status of PROFIT

Beginning around 1992, the microelectronics industry went through a dual transformation in core technology and use (or market). On the technology side, chip size, speed, and versatility took quantum leaps. For example, IBM pioneered copper circuits, RISC based CPU processors, silicon-on-insulator and silicon germanium technologies, and an innovative insulation techniques for copper circuits. The market for microelectronic devices expanded from an initial base in computers to a wide range of products such as cell phone, cars security systems, advanced GPS based trackers, greeting cards, and aids for the handicapped. Microelectronic devices pervade the world. This dual expansion has transformed manufacturing from making a large quantities of just a few parts to varying quantities of numerous parts.

In early 1990s the IBM Corporation made a strategic decision to transition the IBM Microelectronics Division from producing a limited number of parts exclusively for other IBM locations to producing a wide range of products from servers to cell phones for a diverse set of customers. This required transitioning its organization from loosely coupled producing just a few supplied directly to a down stream manufacturing facility to a tightly coupled set of manufacturing facilities To quantify this change in the mid 1980s IBM Microelectronics had about 100 active part numbers with demand. The current number is 6000.

To accomplish this transition the supply chain management applications had to be overhauled to handle the new business environment and the ever increasing complexity of semiconductor manufacturing. A lynchpin of this work was the development of intelligent models to match assets with demand to determine which demands can be met when and provide manufacturing guidelines by the PROFIT (planning resources optimally for IBM technology) team. The team has deployed four core applications: a division best can do (BCD) or central planning engine (CPE) which determines the customer commitments and manufacturing requirements (and is used for what if scenarios), an optimized daily manufacturing resource plan (MRP) which identifies the need date for each lot while making intelligent use of alternative production methods to minimize tardiness, a division available to promise which facilitates fast response to customers placing orders, and demand statement creation function which establishes an integrated demand profile for the entire division. This work has been successful from a technical and business point of view generating a single image data view of the division, pushing the limits of practical large scale optimization, and generating significant improvements in customer service and manufacturing efficiency (Lyon et al. 2001).

IBM Microelectronics Division remains committed to ongoing incremental improvements in SCM decision support. Over the past 18 months emerging business requirements include closing the gap between planning and execution, providing more guidance to execution from planning, reducing the half-life of a plan, emergence of a foundry and design markets, increased in the size of the RTAT (rapid turn around time) market, need for continuous demand management, and improved linkages between SCM components. This represents the first steps in moving from a centralized or big bang approach for planning to incremental modifications. From a decision technology perspective this requires intertwining traditional centralized planning and optimization methodologies with the emerging area of collaboration, intelligent agents, or sense and respond (SaR).

3. Overview of IBM Microelectronics Division

The IBM Microelectronics Division is a leading-edge producer of semiconductor and packaged solutions for the networked world supplying a wide range of customers in market segments such as application specific components, embedded controllers, wireless, microprocessors, memory, and storage. Customers are divided into two major groups: internal and external. Internal refers to other IBM manufacturing facilities that produce such products as mainframes, workstations, storage devices, super computers, and controllers. External refers to all other customers. These customers make such products as workstations, video games, GPS trackers, medical equipment, cellular phones, and many other products.

3.1 Manufacturing Flow

The core manufacturing flow (Figure 14.1) for microelectronic parts is: wafers to devices to modules to cards. The wafer is a round thin piece of silicon that looks similar to a CD. The wafers go through an

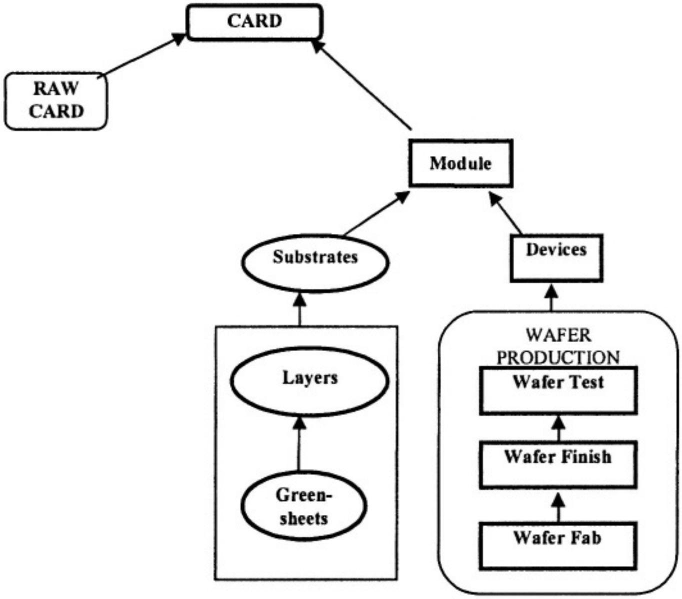


Figure 14.1. Summarized Bill of Material for IBM Microelectronics.

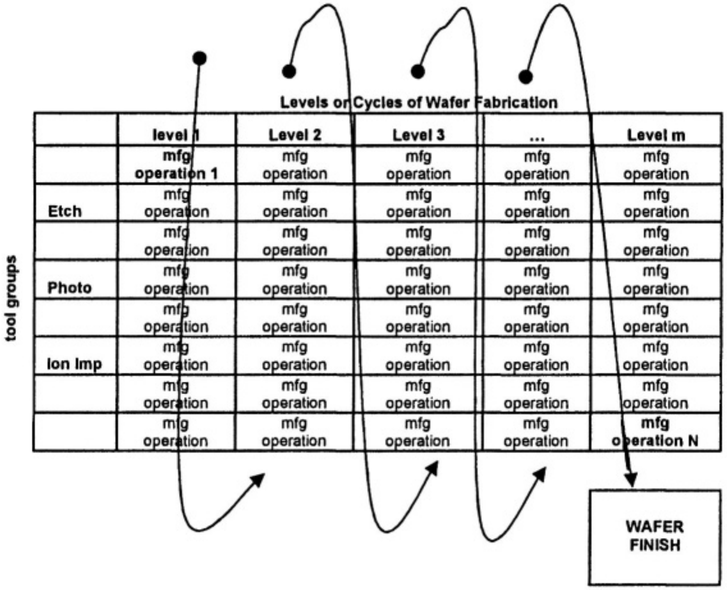
elaborate process that has cycle times (time to complete the task) of 30 to 140 days where thousands of circuits are carefully etched onto it. When the wafer is completed, it is sent to final test and then cut into small individual rectangular shaped parts called devices. Typically cycle times are 5 to 15 few days. The devices are then placed on a substrate and packaged to create a module which takes between 5 to 20 days. Modules are then combined together on a card. Depending on the customer, IBM Microelectronics may ship wafers, devices, modules, or cards. Within each major manufacturing activity there are many individual operations and extensive testing. Operations are grouped into sectors or work centers and sectors are grouped into stages or levels. The actual matching models deployed by the PROFIT team work with dynamically established stages as the core manufacturing activity or decision point. The number of levels range from 4 to 40 depending on the specific model. This chapter will use four levels (wafer, device, module, and card) in most examples.

Wafer production is typically divided into three stages: wafer fabrication, wafer finish, and wafer test. In wafer fabrication the circuits are imbedded into the silicon. In wafer finish the metals to conduct electricity in and out of the chip are placed on the wafer. Each chip is given a full ranges of performance tests in wafer test and appropriately classified.

The process of wafer fabrication begins with a pure, thin, and circular slice or wafer of silicon (which is an insulator). It will eventually become hundreds of chips when the process is complete. Between 200 and 400 complicated operations or steps change the electronic structure of the wafer according to a a very precise plan. The essentials of the circuit or wiring plan for a wafer are a set of masks, one mask for each layer of the chip. The masks are designed by the laboratory. The patterns represent a negative image of the parts of all the transistors and other components to be built into the silicon wafer. Through an oxidation process a protective covering of oxide is grown on the wafer. Next, it is coated with a light-sensitive material called a photoresist. A mask is precisely registered over the wafer and light is projected through the mask onto the wafer, causing the photoresist to harden under clear areas of the mask. The image is developed by washing away the unexposed photoresist. Controlled amounts of such elements as phosphorus or boron are introduced into wafer or ion implantation. In this process dopant atoms are accelerated to a high energy. These atoms strike the wafer and are embedded at various depths depending on their mass and energy. These extra atoms have to squeeze in the best they can with the silicon atoms. Since they typically have one more or one less electron in their outer or valence shell then silicon, they squeeze in by giving up an electron (n-type) or taking on an electron and creating a positive hole (p-type). The wafer is fabricated in layers. Therefore the processes of oxidation, photolithography, and ion implantation are repeated many times until thousands of circuits are built into each wafer.

The flow of wafer fabrication is best represented by a reentrant flow or folded serial line (Figure 14.2). From the wafer's perspective it is produced by following specific sequence of unique operations without any option for variation except for rework loops. From the tooling center perspective each wafer makes approximately many passes or iterations through one of its tools. Each set of tools handles all the activity across all the iterations for that step. Each tool can handle a variety of tasks, and is reconfigured (set up) to handle different wafer types at the specific iterations. For each wafer or part type and iteration step the raw process time is known within a small tolerance (the values range from 15 milliseconds to 20 minutes).

In wafer finish or the metalization phase the wires connecting the transistors are put in. The amount of time and complexity of this manufacturing activity depends on the type of chip being manufactured.



	level 1	Level 2	Level 3	...	Level m
	mfg operation 1	mfg operation	mfg operation	mfg operation	mfg operation
Etch	mfg operation	mfg operation	mfg operation	mfg operation	mfg operation
	mfg operation	mfg operation	mfg operation	mfg operation	mfg operation
Photo	mfg operation	mfg operation	mfg operation	mfg operation	mfg operation
	mfg operation	mfg operation	mfg operation	mfg operation	mfg operation
Ion Imp	mfg operation	mfg operation	mfg operation	mfg operation	mfg operation
	mfg operation	mfg operation	mfg operation	mfg operation	mfg operation
	mfg operation	mfg operation	mfg operation	mfg operation	mfg operation N

Figure 14.2. Folded Serial Line or Re-entrant Flow.

Wafer test is a complex set of activities that determine both the quality of each chip, but determines its performance characteristics resulting in a classification or sorting. Typically, different testers can provide the same function, but with different performance characteristics (speed and accuracy).

In module production different chips are combined with one or more substrates to create the most common finished good. This step differs from wafer fabrication in two critical areas: it is an assembly operation and a large number of substitution and/or alternative production patterns can exist.

3.2 Manufacturing Locations

IBM Microelectronics and its critical suppliers have between 10 and 15 manufacturing facilities in North America, Europe, and the Far East. Typically, a manufacturing facility will specialize in building wafers, wafers and devices, devices and modules, modules, or cards. The allocation of products across manufacturing facilities changes is quite variable and changes regularly. An example of a common product flow is part of the wafer to be made at location A, the wafer completed and device

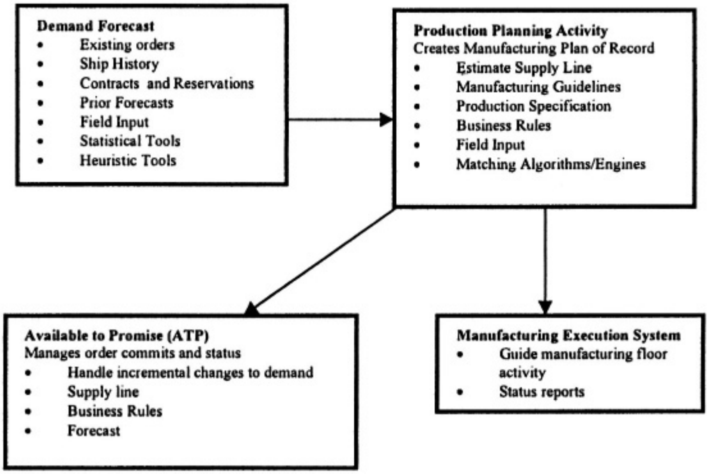

Demand Forecast
• Existing orders
• Ship History
• Contracts and Reservations
• Prior Forecasts
• Field Input
• Statistical Tools
• Heuristic Tools
Production Planning Activity
Creates Manufacturing Plan of Record
• Estimate Supply Line
• Manufacturing Guidelines
• Production Specification
• Business Rules
• Field Input
• Matching Algorithms/Engines
Available to Promise (ATP)
Manages order commits and status
• Handle incremental changes to demand
• Supply line
• Business Rules
• Forecast
Manufacturing Execution System
• Guide manufacturing floor activity
• Status reports

Figure 14.3. Simple Overview of the Supply Chain Process.

created at location B, the device is tested at location C, the module created at location D, the module tested at location C, and then shipped to the customer.

3.3 Challenges - Scope and Scale

The many characteristics of semiconductor manufacturing and the core supply chain management process (Figure 14.3) within IBM Microelectronics that make planning and scheduling a challenge can be divided into two categories: scope and scale. Scope includes complex manufacturing flows, long cycle times, variable cycle times, instability in demands, short product life cycles, and yield (percentage of good parts after manufacturing is completed). Scale includes number of parts, number of customers, and number of manufacturing locations, and very expensive equipment or tools. As a result of this complexity a critical component of effective utilization of manufacturing resources and customer is matching assets with demand (MAWD) intelligently which is the heart of the production planning activity.

MAWD refers to aligning assets with demand in an intelligent manner for a variety of purposes within the supply chain management (SCM) process. The alignment or match occurs across multiple manufacturing facilities within the boundaries established by the manufacturing specifications and process flows and business policies. Assets include, but are not limited to, starts, work in progress (WIP), inventory, purchases, and capacity (manufacturing equipment and manpower). Demands include, but are not limited to, firm orders, forecasted orders, and inventory buffer. The matching must take into account manufacturing or production specifications and business guidelines. Manufacturing specifications and process flows include, but are not limited to, build options, bill of material (BOM), yields, cycle times, anticipated date a unit of WIP will complete a certain stage of manufacturing (called a receipt date), capacity consumed, substitutability of one part for another (substitution), the determination of the actual part type after testing (called binning or sorting), and shipping times. Business guidelines include, but are not limited to, frozen zones (no change can be made on supplies requested), demand priorities, priority tradeoffs, preferred suppliers, and inventory policy. Many of the manufacturing specification and business guideline values will change often during the planning horizon (called date effectivity).

4. PROFIT Core Components, Applications, and Business Value

The PROFIT team built, deployed, and currently supports: a division run which determines the customer commitments and manufacturing requirements, daily manufacturing runs which identify the best use of manufacturing resources to meet the division requirements, a division available to promise which facilitates fast response to customers placing orders, a comprehensive demand statement creation tool, what if planning tools, and eBusiness linkages referred to as customer connect which increasingly provides more value to our customers with the delivery of timely and valuable information.

These application are supported by the following core components.

The best can do (**BCD**) or central planning engine (CPE) SCM activity involves determining how to best meet prioritized demand without violating temporal, asset, or capacity constraints. This application minimizes prioritized demand tardiness in establishing commitments and synchronized targets for each manufacturing location. It creates a projection of what can be produced to meet demand that is a key element of the available to promise (ATP) type of matching.

The optimal manufacturing resource planning **(OMRP)** SCM activity is based on the assumption that a manufacturing facility must meet all demands on time. In theory, the BCD activity has provided each facility targets that are achievable. The OMRP activity provides detailed instructions about what manufacturing activities must be accomplished and when they must be completed. The instructions concern work in progress (WIP), work to be started, purchases, purchase orders, planned substitutions, and shipments. The objective of the optimization portion of this activity is to select production assignments that minimize new starts and starts in negative time. A sister process reviews the optimized assignments and provides an alert manager to possible late orders.

The available to promise (**ATP**) activity enables an organization to dynamically reallocate projected supply in response to incremental changes in the demand statement (new orders arriving, orders being filled, and order changes or cancellations) according to business policy guidelines, identify projected shortfalls with respect to committed orders, and provide real-time order commits and status.

The demand management (**DM**) or demand statement creation activity coordinates demand estimates from different sources such as orders, sales rep forecasts, customer forecasts, internal demand, and marketing forecasts in logical step by step manner.

Figure 14.4 provides an overview of how the PROFIT tools function together. Details about the OMRP and BCD applications can be found in Appendix 2 Additional information about dispatch/manufacturing execution can be found in Appendix 3.

The emergence of eBusiness drove two requirements: delivery of existing information via the web and requests for new information – or at least information customers had not previously request. Examples include lot status, test results, lot history, and BOM path linkages

Currently the PROFIT team supports applications covering order management (order status and ship status), order creation and change, ATP, yield management, WIP coverage, prototype part status, and collaborative forecasting. The order management applications provide summary level data by part order number, full drill down, and email notification. The ATP application enables a user to check availability, examine supply positions, and execute what if modeling of outcomes for various dates and quantities. WIP coverage provides customers visibility of WIP supply against demand by orders and full drill down. Prototype part status enables a customer to follow the progress of prototype parts through the manufacturing process.

The success of this suite of applications has generated substantial improvements in manufacturing productivity and customer responsiveness. The OMRP application reduces required manufacturing starts, provides better guidelines to manufacturing, and establishes better estimates of near term supply. This has reduced the amount of variation of WIP in the line to meet demand reducing cycle time on parts with demand, freed up capacity for other activities, and improved the ability of manufacturing to put energy into the correct activities. The division

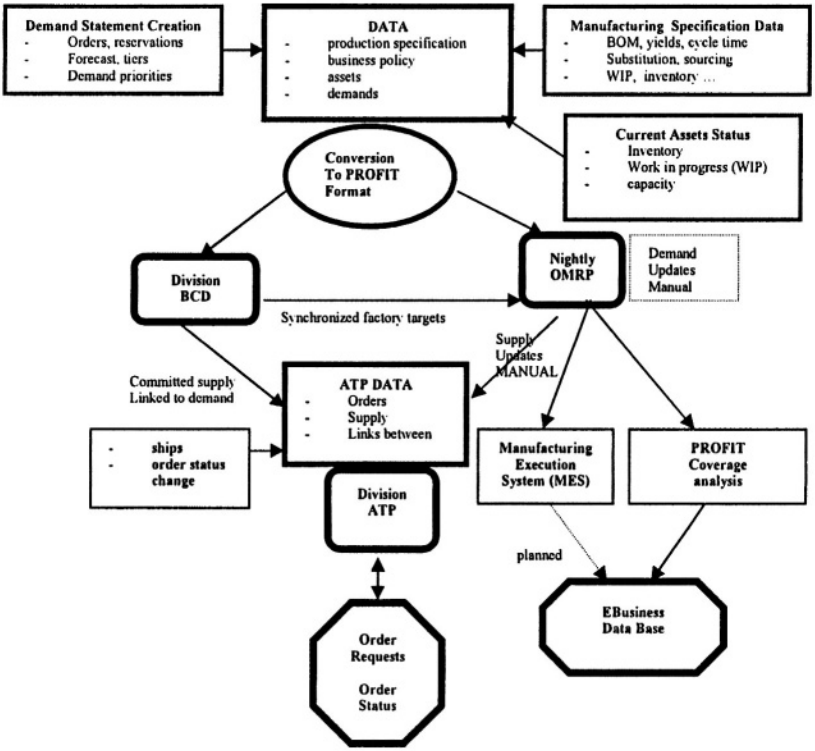


Figure 14.4. Profit Application Flow.

BCD has enabled IBM Microelectronics to have a single source focus of all demand and the supply linked to the demand and provide coordinated manufacturing targets or expectations on each manufacturing facility. ATP, harnessing the coordinated demand and supply picture established by BCD and accurate last minute updates by OMRP, provides rapid responses to request for order requests and order status. The new EDGE applications have provided customers critical information on an on demand basis.

The Microelectronic Division's average order response time (elapsed time between the initial customer order request and IBM's commitment to deliver) has reduced from almost 4 days to 0.6 days as of March 2000. For 70 per cent of the orders, the response time is under 0.3 days and for 90 percent the response time is under 1 day. For many of the orders that come in electronically, the order commit occurs within a few minutes. The ability to commit to the initial customer request date has improved from 10 to 40 percent. On time delivery has increased from 90 percent in 1998 to 97 percent in 1999.

5. Supply Chain Components, Decision Tiers, Centralization vs Collaboration

The key decision points and corresponding decision support applications in the supply chain management process can be classified by the supply chain component supported and its time frame or decision tier. This two dimensional grid provides the fame work for understand where to create decision support enhancements.

5.1 Supply Chain Components

A supply chain management (SCM) process manages the flow of activities from order taking to delivery inside and outside of the organization and consists of business processes, information flows, and decision structures. Figure 14.3 has a very simple high level overview of the supply chain process consisting of four steps: demand creation, production planning, manufacturing execution, and available to promise. The reader will observe many of these core functions are supported in IBM Microelectronics Division by PROFIT applications.

Typically, the first step is **demand statement creation (DSC)** or demand forecasting. The demand statement uses existing orders, history, customer contracts and reservations, field sales force estimates, customer forecasts, and statistical forecasts as key inputs. The result is a demand statement with priorities. In the second step, the **production planning activity (PPA)** matches assets with demand to create an estimated supply line; suggested manufacturing starts, due dates, and priorities; and challenges. The **manufacturing execution system (MES)** receives the suggest starts, due dates, and other guidelines from the PPA and manages the moment by moment activities of the manufacturing floor and creates timely status information. The **available or able to promise (ATP)** module manages incremental changes in the demand statement by matching these changes with the projected supply line. Essentially, it is responsible for meeting yet unspecified orders as they arrive by allocated existing or projected supply.

5.2 Decision Tiers

SCM Decisions in the semiconductor industry typically fall into one of four decision tiers: strategic, tactical, operational, and dispatch (response). The categories are based on the planning horizon, the apparent width of the opportunity window, and the level of precision required in the supporting information.

The first decision tier, **strategic scheduling,** is driven by the time frame or lead time required for business plan, resource acquisition, and new product introduction. This tier can often be viewed in two parts: **very long-term** and **long-term.** Here decision makers are concerned with a set of problems that are three months to seven years into the future. Issues considered include, but are not limited to, what markets will be in, general availability of tooling and workers, major changes in processes, changes in or risk assessment of demand for existing product, required or expected incremental improvements in the production process, lead times for additional tooling, manpower and planning. In the oven example (Appendix 1) very-long-term decisions are made on whether the ovens are necessary to the production process, and if so the characteristics needed in the oven. Long-term decisions are made about how many ovens to buy.

The second tier, **tactical scheduling,** deals with problems the company faces in the next week to six months. Estimates are made of yields, cycle times, and binning percentages. Permissible substitutions are identified. Decisions are made about scheduling starts or releases into the manufacturing line (committing available capacity to new starts). delivery dates are estimated for firm orders, available “outs” by time buckets are estimated for bulk products , and daily going rates for schedule driven product are set. The order/release plan is generated/regenerated. Reschedules are negotiated with or requested by the ultimate customer. In the oven example, average capacity available and required is estimated by part type for use by core matching tools. Typical decision support tools in this tier include production planning, demand forecasting, aggregate capacity analysis, and simulation.

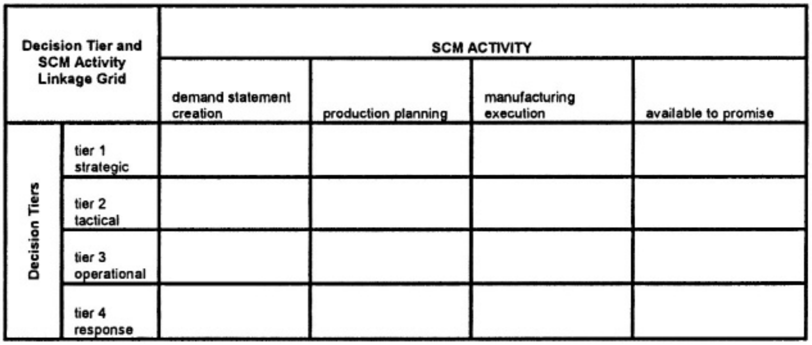
The third tier, **operational scheduling,** deals with the execution and achievement of a weekly plan. Shipments are made. Serviceability levels are measured. Recovery actions are taken. Optimized consumption of capacity and output of product computed. Tools typically use in support of daily activities are material resource planning, decision support, recovery models, prioritization techniques and deterministic forward schedulers. Manufacturing execution systems (MES) are used for floor communications and control. In the oven example, priorities would be placed on each lot arriving at the ovens based on their relevance to current plan or record and detailed capacity estimates for short time horizon would be estimated.

The heart of the production planning type activity in the tactical and operational decision tiers involves matching assets with demand (MAWD) (Appendix 5).

The fourth tier, **real-time response system,** addresses the problems of the next hour to a few weeks by responding to conditions as they emerge in real time and accommodate variances from availability assumed by systems in the plan creation and commitment phases. Within manufacturing the dispatch scheduling (DS) application handles real-time response. Dispatch scheduling decisions concern monitoring and controlling of the actual manufacturing flow or logistics and instructing the operator what to do next to achieve current manufacturing goals. Here decisions are made concerning trade-offs between running test lots for a change in an existing product or a new product and running regular manufacturing lots, lot expiration, prioritizing late lots, positioning preventive maintenance downtime, production of similar products to reduce setup time, down stream needs, simultaneous requests on the same piece of equipment, preferred machines for yield considerations, assigning personnel to machines, covering for absences, and reestablishing steady production flow after a machine has been down. In the oven example the question is which lot (if any) is run next when an oven is free. The goal of most ATP applications is to provide a commit date to a customer order as quickly as possible. Although it may not achieve real-time response its goal is to modify the current match between assets and demand to provide a real-time commit to an order placed by a customer.

5.3 Linkage Between Components and Tiers

There is overlap and interaction between the four decision tiers and four SCM core components and the grid (Figure 14.5) they create. Typ-



Decision Tier and SCM Activity Linkage Grid		SCM ACTIVITY			
		demand statement creation	production planning	manufacturing execution	available to promise
Decision Tiers	tier 1 strategic				
	tier 2 tactical				
	tier 3 operational				
	tier 4 response				

Figure 14.5. Decision Tier and SCM Activity Grid.

ically different groups are responsible for different decisions. Industrial Engineering may have the final say on total manpower, but a building superintendent may do the day-to-day scheduling. Marketing may have the final decision about the demand statement, production planning may determine intermediate target outs, but manufacturing makes the final determination of what is built when. These groups and their associated decision support tools must be coordinated or coupled. A lot only gets processed when the appropriate tool, operator, and raw material are available. An organization achieves this coupling in only one of two ways (Galbraith 1973): slack (extra tooling and manpower, long lead times, limited product variation, excess inventory and people, differential quality, brand loyalty, and so forth) or strong information systems to make effective decisions.

Dating back to the 1950s a centrally controlled intelligent model that ran once every time period (batch or big bang) has been the primary paradigm to deliver decision technology to synchronize or coordinate SCM components and tiers. With each generation of information technology (IT) improvements, improvements in control were achieved either by extending the total area under single control, reducing the execution time of the intelligent models, or some combination of the two. The new century has seen server performance and data exchange capabilities improve to the point where a new paradigm is emerging — collaboration and real-time adjustments.

5.4 Centralization and Collaboration

What is effective centralization? It is essentially the ability to take into consideration all aspects of the decision situation simultaneously and generate one optimal or at least very good solution. To be effective a centralized solution requires a synchronized current view of the entire decision landscape, the ability to deal with complex trade-offs, and fast performance. Even with the best (fastest and most sophisticated) centralized solutions — there are gaps generated by time lags, summarization, performance, triggers, and formulation. By trigger we refer to the event that wakes up and runs the central solution (typically a date). Once a week, once a day, once every 3 days, the central solver executes. The decision to execute is made without any knowledge or monitoring of events since the last execution. Formulation gap refers to the inability to formulate key decision questions in a manner that lends itself to a central solution as opposed to a sequence of negotiations or collaborations.

Why do the gaps matter? Historically, the gaps were closed by small ad hoc applications, smart people, and person to person communication. Just as manufacturing dispatch and scheduling saw in the 1980s (Fordyce and Sullivan 1994), the pace and performance expectations have increased to the point where the current process of filling in the gaps is no longer sufficient to be viable competitor. More efficient processes are required. This is the roll of collaboration which refers to an iterative process that focuses on finding a satisfactory solution and already exists today on an informal basis. The next search process step depends on prior steps and may involve back tracking. Often the step involves negotiating a temporary change in the rules governing the game and typically contingency occurs to handle uncertainty. The market will demand organizations adopt far more sophisticated methods to handle collaboration or perish.

Collaboration does not replace centralization. Without some level of centralized solution, there is no core infrastructure in place to support collaboration. Centralization and collaboration are not competitors, but partners to improve organizational performance. What aspects of decision support within SCM are supported by centralization and which are supported by collaboration can and will evolve as computing technology improves and our experience grows. In fact we might find initial inroads to improve decision support are made with such technology as SaR and later replaced by an improved centralized optimization process.

6. How Do We Get to Collaborative Behavior intertwined with Centralization?

Yes, its hard to start, but so was centralization

6.1 Hypothetical use of Mobile Intelligent Agents for Negotiation

For all of the automation within the current SCM process, most, if not all, of the negotiation that occurs between participating organizations is executed by humans. Humans are good at this, however they have one short coming — speed. As organizations strive for closer and faster collaboration, some of these negotiation tasks will be needed to be handled electronically. The hybrid technology of mobile intelligent agents is emerging as a tool with the promise of handling some of this work successfully. Below is a simple example.

Assume four negotiating agents remote from one another. Agent001 discovers a production problem that might be able to be resolved if agent002 can slip his demand date for a given part. Agent001 users the Directory manager to locate the agent (agent002) with whom he must negotiate. Agent001 dispatches itself to the remote server in which agent002 resides. After deserialization, agent001 engages agent002 with a proposition. Agent002 decides that he cannot slip his requirement (due to his constraints and rules of engagement) unless his customer relinquishes and slips his due date or accepts a substitution. Agent001 refers to the Directory Manager to determine where the appropriate 2nd-level customer agent is (agent003). Agent001 schedules a meeting at the PUMA (PROFIT Universal Mobile Agent) server called NEXXUS. The only message is between agent001 and agent003. Agent002 is serialized and dispatches himself (with state info on the deal in process) to the NEXXUS. Agent001 does the same. When agent003 arrives with state info gathered based on the class and topic of negotiation announced in the meeting notification, the three agents engage in their multi-lateral negotiation in the NEXXUS meeting room. When negotiations are complete, each agent returns (dispatches itself) to its home server and performs any updates appropriate to the negotiation results.

6.2 Near Term Opportunity 1: Demand Statement Creation and Monitoring

There are two areas within the SCM/Decision Tier Grid where we are exploring the use of SaR in a joint project with IBM Research to improve linkage or coupling to improve organizational performance. The first area is demand statement creation and the second is linkage between production planning and dispatch.

The current demand statement creation (DSC) process occurs once a week (sometimes two) and is illustrated in Figure 14.6. On an ongoing basis demand related transactions such as orders, almost orders,

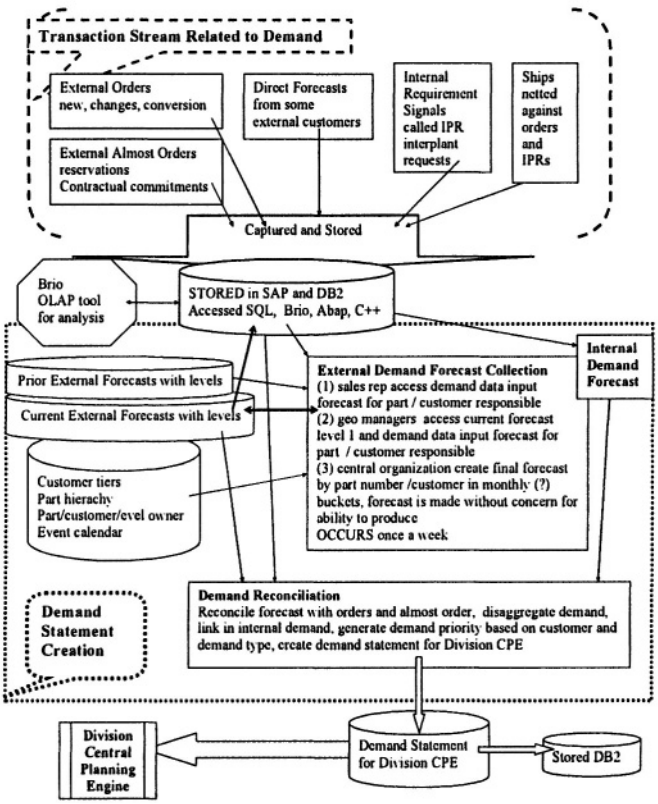

Transaction Stream Related to Demand
External Orders
new, changes, conversion
Direct Forecasts
from some
external customers
Internal
Requirement
Signals
called IPR
interplant
requests
Ships
netted
against
orders
and
IPRs
External Almost Orders
reservations
Contractual commitments
Captured and Stored
Brio
OLAP tool
for analysis
STORED in SAP and DB2
Accessed SQL, Brio, Abap, C++
Internal
Demand
Forecast
Prior External Forecasts with levels
Current External Forecasts with levels
External Demand Forecast Collection
(1) sales rep access demand data input
forecast for part / customer responsible
(2) geo managers access current forecast
level 1 and demand data input forecast for
part / customer responsible
(3) central organization create final forecast
by part number /customer in monthly (?)
buckets, forecast is made without concern for
ability to produce
OCCURS once a week
Customer tiers
Part hierachy
Part/customer/evel owner
Event calendar
Demand
Statement
Creation
Demand Reconciliation
Reconcile forecast with orders and almost order, disaggregate demand,
link in internal demand, generate demand priority based on customer and
demand type, create demand statement for Division CPE
Division
Central
Planning
Engine
Demand Statement
for Division CPE
Stored DB2

Figure 14.6. Simplified View of Core of Current Demand Statement Creation Process.

customer forecasts, and internal demand (interplant requests IPR) come into the organization and are stored. Once a period the DSC process wakes and takes 2-3 days to create new demand statement for use by the division CPE.

In the first step sales representatives are canvassed for their views. They are provided access to the current order book, ship file, and previous forecasts. This group has about a 24 hour period to input via a WEB based application their forecast for the customers and part numbers they have authorization over. In step 2, the geographic managers review the sales rep forecast as well as the other demand related information and generate their forecast. In step 3 the sales and marketing headquarters team reviews all relevant information and generates the

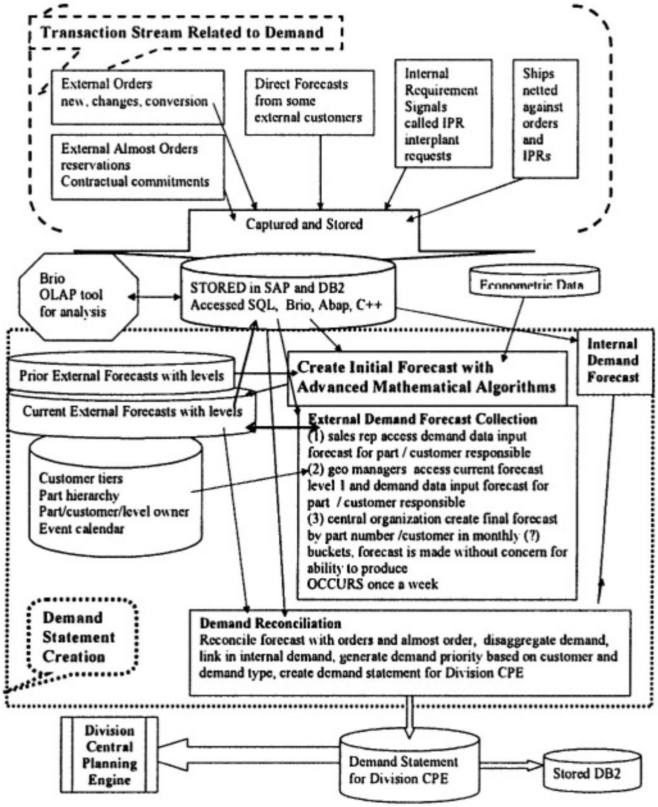

Transaction Stream Related to Demand
External Orders
new, changes, conversion
External Almost Orders
reservations
Contractual commitments
Direct Forecasts
from some
external customers
Internal
Requirement
Signals
called IPR
interplant
requests
Ships
netted
against
orders
and
IPRs
Captured and Stored
Brio
OLAP tool
for analysis
STORED in SAP and DB2
Accessed SQL, Brio, Abap, C++
Econometric Data
Internal
Demand
Forecast
Prior External Forecasts with levels
Create Initial Forecast with
Advanced Mathematical Algorithms
Current External Forecasts with levels
External Demand Forecast Collection
(1) sales rep access demand data input
forecast for part / customer responsible
(2) geo managers access current forecast
level 1 and demand data input forecast for
part / customer responsible
(3) central organization create final forecast
by part number /customer in monthly (?)
buckets, forecast is made without concern for
ability to produce
OCCURS once a week
Customer tiers
Part hierarchy
Part/customer/level owner
Event calendar
Demand
Statement
Creation
Demand Reconciliation
Reconcile forecast with orders and almost order, disaggregate demand,
link in internal demand, generate demand priority based on customer and
demand type, create demand statement for Division CPE
Division
Central
Planning
Engine
Demand Statement
for Division CPE
Stored DB2

Figure 14.7. Near Term Core Demand Statement Creation Process with Advanced Forecasting.

final forecast. This forecast is then sent to demand reconciliation which creates a detailed demand statement by customer, part number, quantity, and demand type by linking the forecast with other information such as the order book and customer reservations. Additionally, the internal demand is integrated with the external demand. This information is then used to drive the division weekly CPE.

Figure 14.7 illustrates the near term enhancement — integration of advanced forecasting algorithms to serve as a starting forecast.

Figures 14.8 and 14.9 illustrate two SaR activities under development. The first are SaR activities to monitor the transaction stream of demand related information. This follows a similar pattern to LMS (Fordyce and Sullivan 1994) making use of a gateway and then incrementally

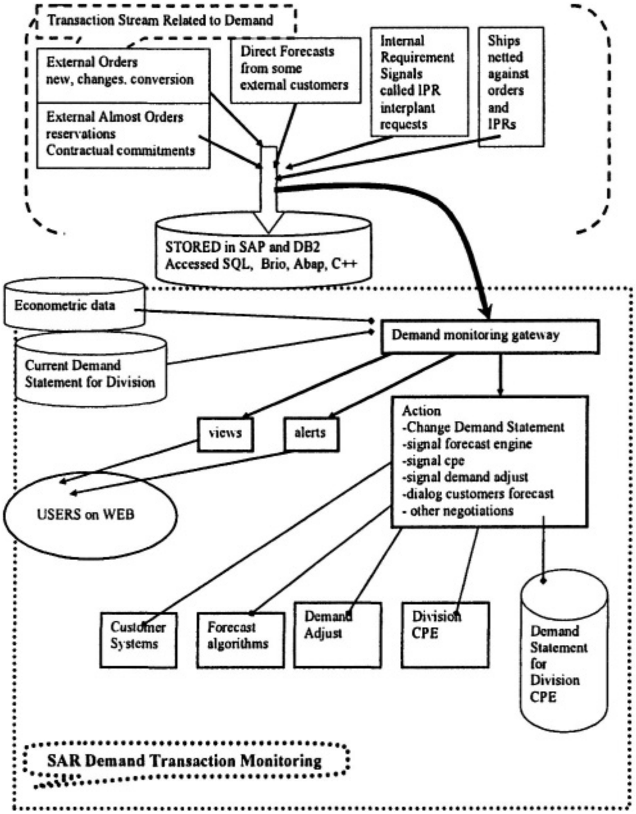

Transaction Stream Related to Demand
External Orders new, changes, conversion
External Almost Orders reservations Contractual commitments
Direct Forecasts from some external customers
Internal Requirement Signals called IPR interplant requests
Ships netted against orders and IPRs
STORED in SAP and DB2 Accessed SQL, Brio, Abap, C++
Econometric data
Current Demand Statement for Division
Demand monitoring gateway
views
alerts
Action
-Change Demand Statement
-signal forecast engine
-signal cpe
-signal demand adjust
-dialog customers forecast
- other negotiations
USERS on WEB
Customer Systems
Forecast algorithms
Demand Adjust
Division CPE
Demand Statement for Division CPE
SAR Demand Transaction Monitoring

Figure 14.8. DSC - Demand Transaction Monitor SaR.

working through the decision support hierarchy from views to alerts to actions. Some of the anticipated actions include altering the demand statement, sending signals to demand adjust, entering negotiations with other incoming forecasts, etc. Figure 14.9 illustrates the use of SaR agents to negotiate between levels during the generation of the weekly forecasting. Currently once one level makes a forecast it has no active role in the creation of the forecast at the next level. This enhancement eliminates this flaw which is particularly critical with the deployment of advanced forecasting methodologies to create a more intelligent starting forecast.

A long term is the creation of a forecasting model that simultaneously considers all factors and creates a plan of record. This would be an ex-

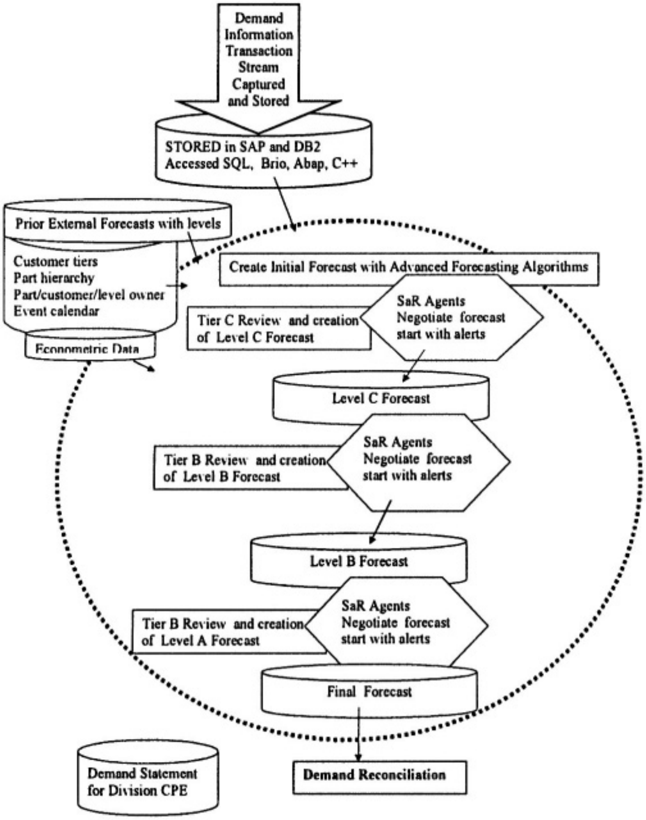

Demand Information Transaction Stream Captured and Stored
STORED in SAP and DB2
Accessed SQL, Brio, Abap, C++
Prior External Forecasts with levels
Customer tiers
Part hierarchy
Part/customer/level owner
Event calendar
Econometric Data
Create Initial Forecast with Advanced Forecasting Algorithms
Tier C Review and creation of Level C Forecast
SaR Agents Negotiate forecast start with alerts
Level C Forecast
Tier B Review and creation of Level B Forecast
SaR Agents Negotiate forecast start with alerts
Level B Forecast
Tier B Review and creation of Level A Forecast
SaR Agents Negotiate forecast start with alerts
Final Forecast
Demand Statement for Division CPE
Demand Reconciliation

Figure 14.9. DSC - DSC Negotiation SaR.

ample of collaboration extending the support provided by centralization only to later be replaced by centralization.

6.3 Near Term Opportunity 2: Linking Execution and Planning

The second area where there are number of related projects to enhance the quality of centralized decision support, build modules to support collaboration, and intertwine these functions to improve organization performance through better decision is linking execution and planning

The old methodology of this linkage is described in Figure 14.10. Once a week a demand statement for the entire division is created, the division CPE (tier 2) executes establishing a supply statement linked to demand and synchronized line outs for each manufacturing facility. The line outs are demand (exploded or dependent) on each facility: part, date, quantity. The overriding theme of this methodology is all of the intelligence is in the plan and none is in the execution. No priority information is provided. It is assumed such techniques as cycle time and yield variability, strict capacitation, and other techniques executed within the CPE insure the lineouts are achievable without a significant loss in potential productivity. These lineouts are sent to each manufacturing location and an MRP runs daily establishing a need date for each lot and aggregate starts required. The MRP does not take into account either capacity constraints or the barrier of negative time. Assuming the CPE has built a flawless plan it doesn't need to. This information is fed to various decision support tools for analysts. During the spring of 2000, an evolution in thinking occurred where the linkages between components needs to be far more dynamic and shared intelligence is required each component.

The starting point for this new paradigm is the replacement of the nightly MRP with a daily CPE (Figure 14.11). Initially the division CPE function will simply run daily on the mfg location information. The daily CPE will send the MES lot sized starts with demand priority and revised demand priority on each WIP lot. This new methodology is represented in Figure 14.11. As you can see from the diagram, there are disconnects between SCM components across decision tiers that are currently handled manually. For example, each night a module called WIP Flush takes in the current location of each lot and estimates when this lot will complete the current stage it is in. This projected receipt is then passed to the daily CPE which then decides how best to use all of the lots. The plan for tomorrow is generated without any linkage to today's plan. The lot may be assigned a new priority or may be slowed down due to capacity restrictions or high priority lots that emerge. However, the

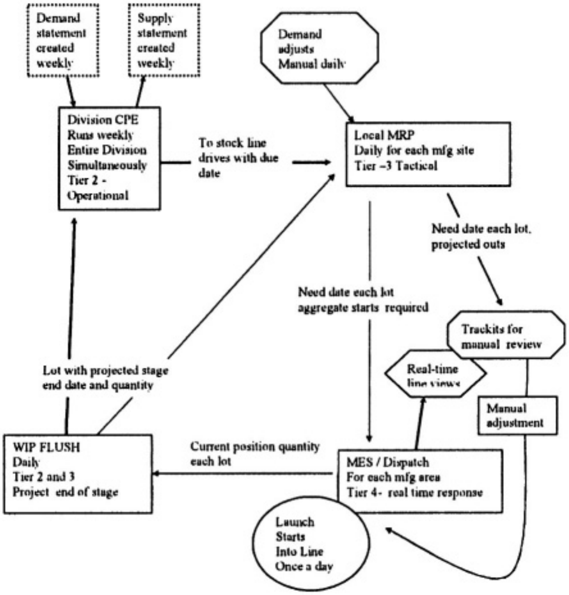

Demand statement created weekly
Supply statement created weekly
Demand adjusts Manual daily
Division CPE Runs weekly Entire Division Simultaneously Tier 2 - Operational
To stock line drives with due date
Local MRP Daily for each mfg site Tier -3 Tactical
Need date each lot, projected outs
Need date each lot aggregate starts required
Trackits for manual review
Real-time line views
Manual adjustment
Lot with projected stage end date and quantity
WIP FLUSH Daily Tier 2 and 3 Project end of stage
Current position quantity each lot
MES / Dispatch For each mfg area Tier 4- real time response
Launch Starts Into Line Once a day

Figure 14.10. Old Production Linkage.

RTD will drive each lot to its original due dates unless the due date is moved in. Other areas where improved coupling can occur are: demand in jeopardy, capacity, and accommodating some demand adjustment.

This was a temporary evolution in thinking. The longer term goal (see Figure 14.11) involves a co-ordinated effort to enhance central support with better BCD tool that runs daily and build a suite of functions called AIIRR (assess, identify, improve, repair, respond) to support collaboration.

The driver is improved ability to respond and adjusting to SCM events a synchronized, timely, and intelligent fashion (soccer team). These events fall into three groups: delta demand, delta supply or asset, and delta business policy. The response has requires actions in one or more of three core areas: DSC (demand statement creation or update), OMAD (optimal match assets with demand), and SATE (synchronization across the enterprise).

The Daily CPE will replace the weekly process, the night nightly OMRP runs, and the site Daily CPE runs. It will create a synchronized plan once a day. The daily division CPE requires improvements in per-

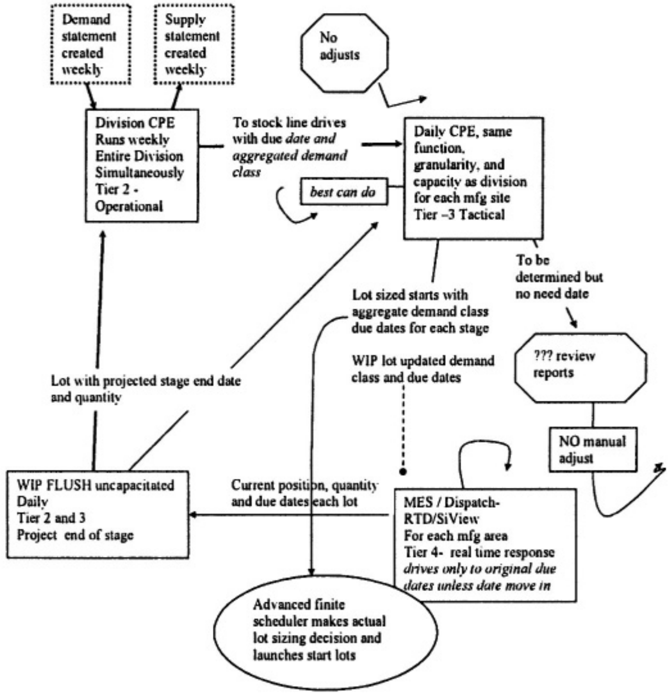

Demand statement created weekly
Supply statement created weekly
No adjusts
Division CPE Runs weekly Entire Division Simultaneously Tier 2 - Operational
To stock line drives with due *date and aggregated demand class*
Daily CPE, same function, granularity, and capacity as division for each mfg site Tier –3 Tactical
best can do
To be determined but no need date
Lot sized starts with aggregate demand class due dates for each stage
WIP lot updated demand class and due dates
??? review reports
NO manual adjust
Lot with projected stage end date and quantity
WIP FLUSH uncapacitated Daily Tier 2 and 3 Project end of stage
Current position, quantity and due dates each lot
MES / Dispatch-RTD/SiView For each mfg area Tier 4- real time response *drives only to original due dates unless date move in*
Advanced finite scheduler makes actual lot sizing decision and launches start lots

Figure 14.11. Near Term Production Signal Linkages with Daily CPE.

formance, increased sensitivity to the solution created the day before, and the ability to move manual modifications in the automated process.

AIIRR (pronounced AIR) is division wide tool to monitor and modify supply picture working in a fully synchronized manner with division wide daily CPE and co-ordinated with customer connect and ATP. Monitor refers to gauging the currently estimated supply against demand. Modification refers to identifying potential opportunities to improve the supply performance outside standard capacity, cycle time, yield and business regulations. The components of AIIRR evolve from passive to pro-active.

The first component is data access. This involves the complete, but easy and rapid access, to all key data from division daily CPE and logically associated data (for example latest demand data and wip location data) from other areas.

The Alert Monitor component involves monitoring data streams from the division daily CPE and past runs for variances of interest to the user and notification of user or person directed by the user through portals, email, or page. It does not identify repair or improvement option

The Repair & Improve Identification component identifies potential options or opportunities to improve or repair a supply picture as requested by the user and puts in place alerts when the user selects this option. It does not evaluate each option.

The Evaluate Intelligent Agent identifies and evaluates each option and reasons from the what if model (see below), puts in place monitor agent, and adds to knowledge base dynamically based on dialogue with user.

The Collaborative component opens dialogues with other participants to see if a recommended solution is acceptable to them. Other participants would include manufacturing, customers, management, marketing to change demand or demand priorities

The Rapid Response What IF model (links back to division daily CPE team) is sort of a short almost interactive CPE. It handles a what if scenario evaluation to gauge impact of change and , enables agents to reason beyond just rules and manipulation demand pegging changes

7. Incremental Matching- The Holy Grail

The current paradigm that dominates matching assets with demand (MAWD) in tiers 2 (tactical) and 3 (operational) of the SCM process is the “big bang” approach. Manufacturing status data (WIP, capacity), manufacturing specification data (yields, cycle times, bill of material flows), business policy information (buffers, lead times, priorities, build to forecast percentages), and demand information is gathered up in bulk every cycle (daily, weekly, or monthly) to feed to a large MAWD engine which either does a BCD match or an MRP match. At this time a new game plan is established, an estimated supply line is created and guidelines are sent to the next tier (tactical sends guidelines to operational, operational sends guidelines to dispatch).

To date improvements in the matching process have focused on doing the big bang match more often (for example moving the division match from monthly to weekly) and using less summary and more detail data (for example running at a part number level and handling lot sizing). The fundamental paradigm remained the same. To date the paradigm has worked adequately. However the fundamental paradigm contains some inboard disconnects: (1) The data used by the MAWD engines is not as current or detailed as the data in the tier below. For example a typical prestep before executing a MAWD engine is to project WIP "to stock". A deterministic algorithm using average cycle times estimates the date at which the lot will arrive at the next manufacturing activity point handled by the matching engine. (2) The plan starts to become obsolete immediately after being created. For example in a weekly division run the project supply plan is on average 3.5 days old. (3) There is no referential integrity to the last plan. (4) To have the MAWD engine run fast enough to support interactive what if analysis, the model has to run with summary level data (parts and time buckets). (5) Deep dynamic ATP calculations with short response time is at best very difficult to obtain. As Galbraith (1973) observed limitations in information flow are handled by "slack".

To date the "big bang" paradigm has served semiconductor firms well. However with the emergence of eBusiness and the subsequent generation of new expectations this paradigm will need to be replaced by incremental matching – or a unified field theory of SCM activity/decision tiers grid.

Conceptually, with incremental matching there will be a global state array with the status of each WIP element, projected start, capacity allocation, and manufacturing guidelines. A set of bond values will associate each of these elements with other WIP elements and demand elements. Mobile intelligent agents data harvesters (MIADH) will reside in each critical source data watching each transaction into the data source. Using its intelligence it will forward pertinent changes to the global state array matching (GSAM) engine. Based on the submitted transaction the GSAM engine can dispatch other agents to (1) gather additional information, (2) negotiate, and (3) incrementally modify the existing plan and guidelines. Although much broader in scope than the gateway in LMS, the core concepts are similar.

8. Conclusion

"It is a new kind of entity. It's a COMMUNITY INTELLIGENCE, born from the collective wisdom of various disciplines, experiences, and points of view, which dynamically disseminates the new intelligence around the same community that engendered it, solving problems that are 'too tough for us humans to figure out" ' (Feigenbaum, pp. 63-64).

This quote could have come from one of the many eBusiness press releases made over the past months. In fact the quote is from 1988 and appeared the book **The Rise of the Expert Company.** This book reviewed in depth critical applications in large corporations that used the then "new" technology of expert systems to improve decision making in reasonably narrow areas when compared to the breadth of the SCM Activity/Decision Tier grid.

There is no reason information and decision technology can not be interweaved to achieve this goal for supply chain management. The key is the right intermingling of collaboration with centralization and evolutionary approach to development. As with any major exploration activity there will be uncertainly and temporary setbacks. The right paths can be found the goals can be achieved.

Appendix 1: Oven Dispatch Example to Illustrate Decision Tiers

Wafers move around in groups of 25 called a lot. All wafers in the lot are the same type. Each lot must pass through the oven operation 10 times. Each Oven set is composed of four ovens or tubes and 1 robot to load and unload the oven. It takes about 10 minutes to load or unload an oven. The process time in the oven depends on the iteration. We will assume one lot to an oven at a time. Before a wafer enters into the oven it must be coated. The coating process takes 20 minutes. The coating expires in four hours. If the coating expires the wafer must be stripped, cleaned, recoated. This process takes 4 hours and often generates yield losses.

Appendix 2: Technical Overview of the PROFIT BCD and OMRP Matching Applications

The heart of the OMRP and BCD applications is moving work units (WIP or starts) either forward to project completed parts or backwards to determine starts required across the bill-of-material (BOM) chain following the appropriate manufacturing movement rules such as cycle time, yield, capacity, and product structure. We typically use implosion to estimate what finished goods will be available to meet demand and explosion to estimate required what starts are needed at what due dates to insure meeting the existing demand on time.

To review implosion and explosion we reference Figure 14.12, which represents a simple production or bill of material (BOM) flow. The first manufacturing activity is the production of Wafer_2. This manufacturing activity has a cycle time of 30 days.

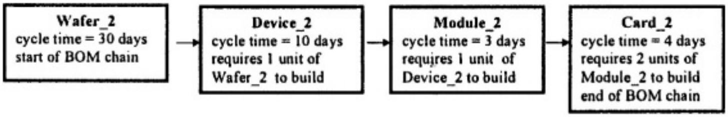


Figure 14.12. Simple BOM Chain to Illustrate Explosion and Implosion. The first manufacturing activity is wafer production. A wafer becomes a device which becomes a module which becomes a card. In implosion, WIP is projected forward to its final good form. If four device units be will be on day 10, these four devices are projected to be 4 modules on day 13 and 2 cards on day 17. In explosion a demand projected backward to determine the start needed to make this demand. If one card is required on day 20, then two modules are required on day 16, two devices on day 13, and two wafers on day 3.

That is, it takes on average 30 days to take a raw wafer and create a completed wafer with the part id Wafer_2. The second activity is device production. To create one unit of Device_2 requires 10 days of cycle time and the consumption of one unit of Wafer_2. Module 2 consumes one unit of Device 2 and takes 3 days to produce. Card 2 consumes two units of Module 2 and takes 4 days to produce.

Referencing Figure 14.12, implosion can be illustrated with the following example. Manufacturing estimates four units of device 2 will be available or completed on day 10. This is called a projected receipt. If manufacturing immediately uses these four units to produce Module_2, then on day 13 (10 + Module_2 cycle time = 10 + 3 = 13) four units of Module_2 will be completed. Continuing the projection process, the four units of Module_2 are immediately used to create two units of Card_2 which will be available on day 17 (13 + cycle time for Card_2 = 13 + 4 = 17). The implosion process enables manufacturing to estimate the future supply of finished goods.

Again referencing Figure 14.12, explosion can be illustrated with an example. To meet demand for one unit of Card_2 on day 20, the plant must have two completed units of Module_2 available on day 16 (20 minus the cycle time for Card_2 = 20 - 4 = 16). This generates an exploded demand of two units of module 3 with a due date of day 16. To continue the explosion process, to produce the two units of Module_2, the plant must have units of Device_2 available on day 13 (16 minus the cycle time for Module_2 = 16 - 3 = 13). Next, The device demand is exploded creating a demand for units of Wafer 2 on day 3 (13 - 10). This exploded information creates the guidelines for manufacturing to meet existing demand. For example the device department must start production of two units of device 2 no later than day 3 to meet the demand for one unit of Card_2 on day 20.

Optimized Material Resource Planning Application

The optimized material resource planning (OMRP) application provides detailed guidelines to manufacturing and a detailed estimate of supply. It contains such traditional MRP features as lot-level details, lot sizing, daily time buckets, and the ability to handle partial day cycle times. In addition, it has the ability to optimally allocate an asset when there are competing demands for this asset. It consists of three core components: binning material resource planning (BMRP) module, alternative BOM material resource planning (AMRP) module, and the partitioning module (PARTITIONER).

OMRP was designed to handle five core decision technology challenges: (1) simple binning with downgrade substitution, (2) alternative production processes for the same part (alternative BOM structure), complex binning, and general substitution, (3) granularity at the individual manufacturing lots with cycle times that have partial days, (4) handling production specification information like yields, cycle times, available and capacity that change regularly during the planning horizon (date effectivity), and (5) determining the optimal match between assets and demand in a reasonable amount of time (performance).

Simple Binning and BMRP

Within the production of devices simple binning (Figure 14.13) with down grade substitution is a common manufacturing characteristic. Binning or sorting refers to the characteristic of assigning a part a specific identity only after testing it. After the wafer is completed and tested there is a 50% chance it will be classified as device A, 30% chance it will be classified as device B, and 20% chance it will be classified as device C. These values are called the binning percentages and the devices are referred to as co-products. Additionally, device A can be used (substituted) to meet demand for devices B and C.; device B can be used to meet demand for device C. This is called down-grade substitution.

The challenge is to make optimal use of co-products and substitution to avoid overstating the required starts need to meet demand. If the demand for devices is 30 for Device A, 40 for Device B, and 30 for Device C, the challenge is to determine the minimum number of wafers that must be produced to meet all of the demand. If we optimally account for co-products and substitutions the minimum number of wafers required to meet this demand is 100. Starting 100 wafers creates 50 of A, 30 of B, and 20 of C. The extra 20 of A are used to cover the shortfall of 10 of B and 10 of A.

Other factors complicating the determination of the minimum number of starts required to meet demand in simple binning production structures: demands for devices are spread throughout planning horizon, existing inventory, projected completion of WIP, and binning percentages and allowable substitutions that change during the planning horizon (date effectivity).

The PROFIT team developed the BMRP module of OMRP to handle simple demand which includes simple binning. Parts that are classified as **simple demand** have no alternative BOM structures, general substitution, or complex binning in their production path. These demands may have one or more simple binning in their production path. The BMRP module is similar to a traditional MRP explosion algorithm except at simple binning points within the BOM chain. At these manufacturing activities a small LP is invoked to determine the optimal required starts. The PROFIT team developed an LP formulation which calculates the minimum number of required starts each day spread across a planning horizon for a specific binning activity, algorithms and data structures to dynamic identify of each instance of simple binning

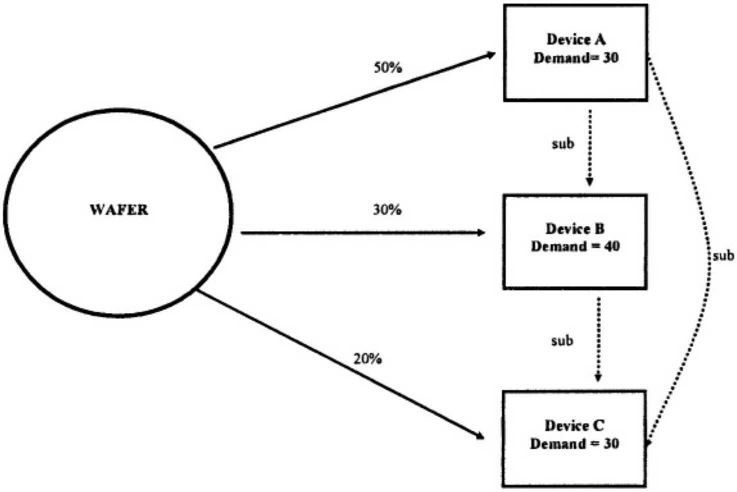


Figure 14.13. Example of Simple Binning with Substitution. Binning or sorting refers to the characteristic of assigning a part a specific identity only after testing it. After the wafer is completed and tested there is a 50% chance it will be classified as device A, 30% chance it will be classified as device B, and 20% chance it will be classified as device C. These values are called the binning percentages and the devices are referred to as co-products. Additionally, device A can be used (substituted) to meet demand for devices B and C.; device B can be used to meet demand for device C. This is called down-grade substitution. If the demand for devices is 30 for Device A, 40 for Device B, and 30 for Device C, the challenge is to determine the minimum number of wafers that must be produced to meet all of the demand. If we optimally account for co-products and substitutions the minimum number of wafers required to meet this demand is 100. Starting 100 wafers creates 50 of A, 30 of B, and 20 of C. The extra 20 of A are used to cover the shortfall of 10 of B and 10 of A.

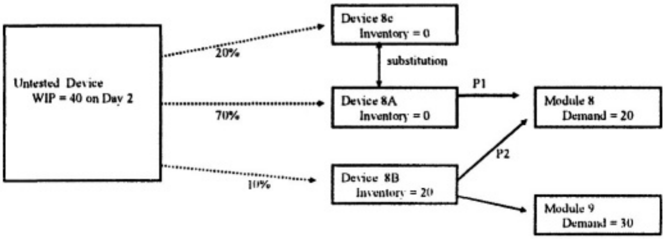


Figure 14.14. Example complex BOM chain with binning, alternative production paths, inventory, and WIP.In this BOM chain 40 units of WIP at untested device will be available on day 2. The binning percentages for devices 8C, 8A, and 8B are 20%, 70%, and 10% respectively. Module 8 can be made by two different manufacturing processes. Process 1 consumes Device 8A and Process 2 consumes Device 8B. Device 8C can be substituted for Device 8A. Module 9 can only be built with Device 8B. There is 20 units of demand on Module 8 and 30 units on Module 9. The objective for an intelligent explosion algorithm is allocating demand for Modules 8 and 9 back across production to make best use of existing inventory and WIP, to minimize new starts, and meet other relevant guidelines.

within a traditional MRP explosion process, and linkages to dynamically execute the binning LP model as needed. Details of the binning LP formulation are provided in Appendix 1.

Alternative BOM structures, general substitution, and complex binning

Within the production of modules an increasingly common manufacturing characteristic is alternative production options (called alternative bill of material (BOM) structures), general substitution, and complex binning. Complex binning refers to a situation where one binning activity immediately invokes another or substitutions are permitted across binning activities. When alternative methods are available to produce a part the tool to handle explosion must select between one of two or more paths in propagating demand back through the BOM structure. With alternatives comes the requirement for search to find the best alternative.

To explain these decision challenges we will use Figure 14.14. Two processes (P1 and P2) can be used to build Module_8. The P1 process consumes Device_8A and the P2 process consumes Device_8B. The explosion engine must determine how to explode demand for Module_8 back to the device level? Half to P1 and half to P2, 2/3 to P1 and 1/3 to P2, all to P1? The objective is to divide the demand for Module 8 across P1 and P2 to make best use of existing inventory and WIP, to minimize new starts, and meet other relevant guidelines (for example sharing percentages). Determining the best result requires an extended search through the entire BOM structure.

The demand for Module_8 is 20 units. Twenty units of Device_8B are in inventory, which can be used to make Module_8 with process P2. Also Device_8A can be used to build Module_8 with process P1. There is no current inventory at Device_8A. Most of the search engines in heuristics that guide explosion through alternative BOM structures would explode the 20 units of demand for Module_8 down the P2 leg or process. However, a broader search would uncover available options at untested device and avoid the conflict for Device_8B between Module_8 and Module_9.

There are 30 units of demand for Module_9 and Module_9 can only be made from Device_8B. Are there other options to meet the demand for Module_8? There are 40 units of projected WIP at the untested device, after binning or sorting 8 will become Device_8C, 28 will become Device_8A, and 4 will become Device_8C. Since Device_8C can be generally substituted for Device_8B, there are 36 (8+28) future devices that can be used to produce Module_8, but not Module_9. It is probably not optimal to explode the demand for Module 8 down the P2 leg.

When alternative BOM structures and general substitutions are part of the manufacturing process, explosion is challenging. Because so many factors are relevant (inventory, WIP, demand, binning percentages, permissible substitutions) at multiple bill of material levels and because these factors vary across time, we thought that linear programming was the best way to solve such problems.

Demand on part numbers which have alternative BOM structures, general substitution, or complex binning as part of their production process are called **complex demands.** To optimize the explosion process decisions on complex demands the PROFIT team developed the large LP. The entire explosion process is represented in the large LP equations and the implementation is completely data driven. The core decision variable of the large LP is the quantity of starts at each manufacturing activity during each time bucket. The objective is to minimize a weighted average tardiness in meeting demand on the date requested. Each demand is placed in one of multiple demand classes. The material balance equations represent the entire BOM chain and asset (starts and WIP) movement in all of their complexity. These equations accommodate binning, alternative BOM structures, substitution, different shop calendars, date effectivity, capacity, and sourcing. They insure temporal feasibility. The large LP supports variable time buckets. The large LP can be used as an MRP tool or a BCD tool with only minor adjustments.

Lot Sizing, Lot Identity, and Daily Granularity

The large LP does a superb job optimizing across the complexities created by binning, substitution, and alternative BOM structures. However, it lacks three key features of traditional MRP explosion algorithms: lot sizing, maintaining lot identity, and daily granularity. An application to provide daily manufacturing guidelines without these three features is worthless.

The key challenge in traditional explosion is allocating exploded demand among the alternative paths generated when binning, substitution, or alternative BOM structures occur in the production process. The key challenge in linear programming is maintaining detail granularity. Traditionally, linear programming models aggregate production information into time buckets and part buckets. As a result critical lot level and daily detail information is lost. The PROFIT team developed the advanced material resource planning (AMRP) module wich contains a unique method to interweave these two decision technologies to gain the best of each. AMRP invokes, when necessary, the large LP at each low level code iteration of a traditional MRP explosion process to identify how to optimally allocate substitutions, select from alternative BOM paths, and calculate starts at binning. Low level code refers to assigning each manufacturing process a number indicating its level in the manufacturing process. Manufacturing activities which produce parts that are used to meet customers demand and are never consumed by any other manufacturing activity are assigned a low level code value of 1. In the example in Figure 14.12 the manufacturing activity Card_2 has a low level code of 1. Manufacturing activities which produce parts that are consumed only by manufacturing activities assigned a low level code of 1 are low level code 2 parts. Low level code 2 activities may produce parts which are shipped directly to customers in some cases, but there is at least one instance where the part is consumed by a low level code 1 manufacturing activity. In the example in Figure 14.12 the manufacturing activity Module_2 has a low level code of 2.

We will explain how PROFIT interweaves traditional MRP explosion and the large LP in AMRP using the example portrayed in Figure 14.15. First AMRP converts all alternative BOM structures to equivalent substitution structures. It then runs the large LP on all low level codes (all parts from module through wafer) posting the optimal use of substitutions as receipts (expected completion date of a part). AMRP then invokes the MRP explosion engine and explodes demands from low level code 1 (LLC-1) to LLC-2 using the posted receipts to guide its allocations across alternative BOM structures and its consumption of the existing substitutable supply. In the example in Figure 14.15, the large LP will instruct the traditional MRP explosion how to allocate exploded demand for Modules 1 and 5 between the two alternative processes available for each (P1 and P2). At LLC-2, there are no complex substitutions or alternate processes for these raw modules. Consequently, traditional MRP logic is sufficient and there is no value in running the more time consuming large LP at LLC-2. The LLC-2 demand is the exploded demand from the low level code 1 explosion plus any independent demand for LLC-2 parts. MRP logic then explodes this demand to LLC 3. At LLC-3, complex substitutions are permissible, so it is necessary to run the large LP from LLC-3 (devices, wafers, and common wafer) to optimize this level. The large LP will provide guidelines at LLC-3 on how to use substitutions between devices 1A2 and 1B2 optimally. Subsequent LLC iterations are not complex, so the large LP is not used again.

Performance

To resolve the performance challenge, the PROFIT team deployed deploying two dynamic partitioning strategies and implemented parallel computation on an IBM super computer. In the first partition, the PARTITIONER module divides demands and their BOM parts and structures into two groups: complex and simple. The complex group is solved with AMRP, and the simple group is solved with BMRP. In the second partition, this module divides the complex group into logically independent groups that can be solved in parallel by separate LP runs.

Division Wide Best Can Do (BCD) Application

OMRP and BCD support different, but related, business functions. Both need the ability to move WIP and starts across the BOM chain and search for optimal matches. The BCD application determines which demand can be met when. Its focus is on the overall allocation of manufacturing assets among competing demands. The OMRP assumes the demand it is given by the BCD application can be met and focuses on the optimal use of existing assets and providing detailed guidelines to manufacturing. The BCD does not need to accommodate such details as lot sizing, lot identity, and daily granularity. Daily granularity means identifying the arrival and departure of an asset from a manufacturing activity on a specific date (and time if needed). The BCD is able to operate with time buckets. For example, identifying the arrival or departure of an asset to a specific week instead of a specific day. The core decision challenge to the BCD application is intelligently allocating manufacturing assets to a prioritized demand spread across a planning horizon directed toward minimizing total tardiness when different demands have different priorities and therefore different tardiness weights.

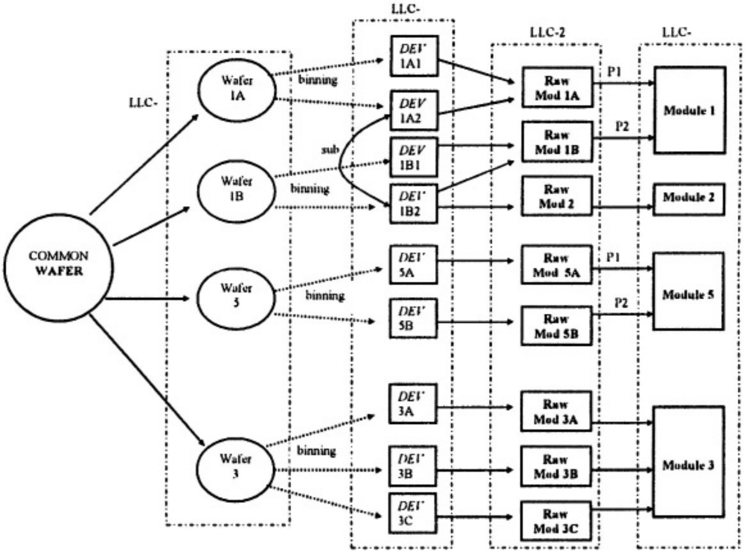
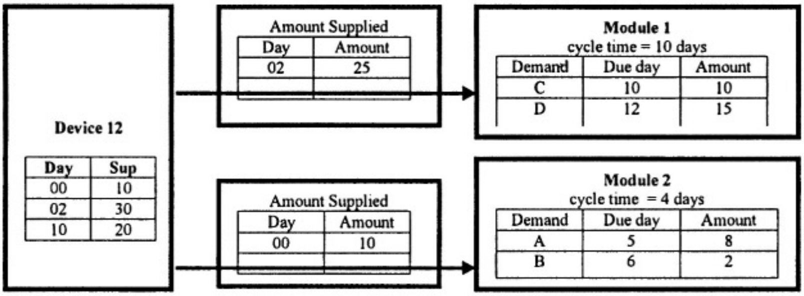


Figure 14.15. Complex BOM Structure with manufacturing levels. A complex BOM flow is presented which has binning, alternative BOM, and substitution. The low level code for each group of parts is displayed. The AMRP interweaves LP and a traditional MRP explosion algorithm by running the large LP at each level to select the optimal use of substitutions and alternative BOM paths. This information is then used by the explosion algorithm in a traditional one level explosion.



Device 12

Day	Sup
00	10
02	30
10	20

Amount Supplied

Day	Amount
02	25

Module 1
cycle time = 10 days

Demand	Due day	Amount
C	10	10
D	12	15

Amount Supplied

Day	Amount
00	10

Module 2
cycle time = 4 days

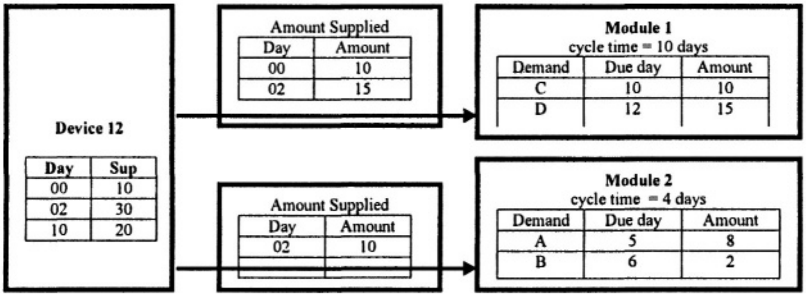
Demand	Due day	Amount
A	5	8
B	6	2

Figure 14.16. On time delivery example - option 1. Module 1 and Module 2 are both made from Device 12. The cycle time is 10 days for Module 1 and 4 days for Module 2. The demand for Module 1 is 10 units on day 10 and 15 units on day 12. The demand for Module 2 is 8 units on day 5 and 2 two units on day 6. In this solution 10 units of Device 12 are allocated to Module 2 on day 0 and 25 units of Device 12 are allocated to Module 1 on day 2. With this allocation Demand A is met 1 day early, Demand B is met 2 days early, Demand C is met 2 days late, and demand D is met on time.

To explain the challenge the BCD application faces we will use figures 05 and 06. Module 1 and Module 2 are both made from Device 12. Demand for Module 1 is ten units on day 10 (demand C) and fifteen units on day 12 (demand D). The cycle time to build Module 1 is 10 days. The demand for Module 2 is eight units on day 5 (demand A) and two units on day 6 (demand B). The projected supply for Device 12 is ten units on hand now (0 days), thirty units on day 2, and twenty units on day 10. The problem is how to allocate the anticipated supply of Device 12 parts between Module 1 and Module 2.

Option 1 (Figure 14.16) might be: (1) immediately allocate eight of the ten units of Device 12 on hand to meet demand A (eight units of Module 2 on day 5) one day early (on day 4=0+4); (2) to immediately allocate the remaining two units of Device 12 on hand to meet demand B (two units of Module 2 on day 6) two days early (on day 4=0+4); (3) on day 2, to allocate 10 units of the projected supply of 30 units of Device 12 to demand C (10 units of Module 1 on day 10) two days late on day 12 (12=2+10); (4) on day 2, to allocate 15 units of the projected supply of 30 units of Device 12 to demand D (15 units of Module 1 on day 12) on time (12=2+10). The score card for this solution is demand A early, demand B early, demand C late by two days, and demand D on time.

A second option is illustrated in Figure 14.17. By searching we can find alternative solutions to the problem. The key questions for the search engine is when should truncate search and how to evaluate the relative merit of each alternative. For example, if the priority on demand A (eight units of Module 2 on day 5) was a much higher then the priority on demand C (10 units of Module 1 on day 10), option 1 would probably be preferred over option 2. The PROFIT team relied on three core components in designing the BCD application: (1) the large LP to handle complex demand, (2) the IMEX BCD heuristic to handle simple demand, and (3) a dynamic partitioning algorithm to split demand between the two BCD solvers and to create parallel computational opportunities.



Day	Sup
00	10
02	30
10	20

Day	Amount
00	10
02	15

Day	Amount
02	10

Demand	Due day	Amount
C	10	10
D	12	15

Demand	Due day	Amount
A	5	8
B	6	2

Figure 14.17. On time delivery example - option 2. Module 1 and Module 2 are both made from Device 12. The cycle time is 10 days for Module 1 and 4 days for Module 2. The demand for Module 1 is 10 units on day 10 and 15 units on day 12. The demand for Module 2 is 8 units on day 5 and 2 units on day 6. In this solution 10 units of Device 12 are allocated to Module 1 on day 0, 15 units of Device 12 are allocated to Module 1 on day 2, and 10 units of Device 12 are allocated to Module 2 on day 2. With this allocation Demand A is met 1 day late, Demand B is met on time, Demand C is met on time, and demand D is met on time.

The IMEX BCD heuristic is a high speed heuristic relying on an explode/implode paradigm to match assets with demand in the best way possible. This MRP based heuristic has three parts.

We use a special variation of the BMRP to explode demand across the entire BOM chain. During the explosion portion, the heuristic optimizes simple binning situations and analyzes key resultants against constraints to establish guidelines for the subsequent implosion. After completing the explosion, the application posts three flies: capacity required (at each manufacturing point where capacity is measured), starts required, and need dates for all projected receipts (need date for anticipated completion of WIP). The starts file lists demand priorities for the entries based on the original demands. In the second step, a user or another program can modify the starts file, the projected receipts file, and the capacity available file. Third, an implosion based heuristic creates a projected supply of finished goods and estimated commit dates for demands that meets all constraints (temporal, asset based, and business policy). IMEX runs substantially faster than the large LP for BCD, but is not as intelligent. IMEX's major weakness is handling complex demands.

Appendix 3: Detail Description of Small or Binning Linear Programming Model

The purpose of this small LP is to determine the minimum production of the binned part required so that the demand of all output parts is satisfied on time. The output parts $(j = 1 \ldots J)$ are the parts that result when the binned part is produced. Usually, these output parts are the same as the parts with demand $(k = 1 \ldots K)$. In Figure 14.14, for instance, the wafer is the binned part and devices A, B, and C are the output parts, each of which has demand.

Definition of Subscripts:

$j =$ output part number which results from the binning process $(j = 1 \ldots J)$

k = part number with demand ($k = 1 \ldots K$)
t = time period ($t = 1 \ldots T$)

Decision Variables:
$P(t)$ = production of the binned part in period t
$S(j,k,t)$ = quantity of output part j used to satisfy demand of part k in period t (note: j may equal k)
$I(j,t)$ = inventory of output product j at the end of period t (which is a function of P and S)
$Z(k,t)$ = unsatisfied demand of part k during period t

Parameters:
$D(k,t)$ = demand for part k in period t
$I(j,0)$ = inventory of output part j available at the beginning of the horizon ($t = 0$)
$R(j,t)$ = fixed receipts of part j in period t
$B(j,t)$ = binning percentage, i.e., percentage of output part j resulting per piece of production of the binned part in period t (includes yield as well as binning distribution)
$p_1, p_2, \ldots, p_5$ = Cost parameters

Objective Function:

$$\text{Minimize} \sum_t \left[P(t) + \sum_j \left[p_1 p_2 Z(j,t) + p_3 I(j,t) + \sum_k \left[p_4 S(j,k,t) \right] \right] p_5 t \right]$$

Constraints:

$$\begin{aligned} I(j,t) &= I(j,t-1) + B(j,t) * P(t) + R(j,t) - \sum_k S(j,k,t) \\ D(k,t) &= Z(k,t) + \sum_j S(j,k,t) \\ & \quad I(j,t), P(t), Z(j,t), S(j,k,t) \geq 0, \forall j,t,k. \end{aligned}$$

Comment: Typically, the Z variables will be zero since the LP model is aiming to satisfy all demand on time. However, the Z variables are necessary to prevent infeasibilities which may otherwise result from bad input data.

Appendix 4: Characteristics of Dispatch in Semiconductor - Overview of LMS Architecture

Complexity in scheduling the actual manufacturing activity within semiconductor facilities is due to (Fowler) such factors as unreliable equipment, batching, reentrant flows, rework, yield loss, hot lots, combination of production, engineering and R&D lots, fluctuation of demand priorities, and varying product mix and start rates. There are a few characteristics which reduce complexity (Fowler): (1) Except for rework, most of the flow in a fab is deterministic. (2) The processing time per wafer or per lot or per batch is very nearly deterministic, so that once processing begins, we can get a very good prediction of when the processing will end. (3) The shop floor control systems in place in current wafer fabs provide much of the information we need in order to make good decisions.

Logistics Management System (LMS) was built in the middle 1980s to support this tier 4 (dispatch) SCM activity. LMS is a real-time transaction-based system using various decision technologies to serve as a dispatcher, monitoring and controlling the manufacturing flow of semiconductor facilities. It coordinates the actions and decisions of several logically isolated participants in a serially dependent system of activities. Therefore it balances the requirements of several goals (cycle time, output, on time delivery or serviceability, and inventory management) which compete for the same resource, exploits emerging opportunities on the manufacturing floor, and reduces the distortion from unplanned events. Support in LMS comes in two flavors passive (or decision support) and pro-active (or intervention). In the decision support mode LMS passively waits for the user to make a request for information. In the intervention mode LMS monitors the transaction stream and actively uses its knowledge bases and models to issue alerts and recommend what actions to take next.

To accomplish this task LMS captures and stores in real-time all manufacturing transactions, and maintains required knowledge bases and models. LMS provides the dispatch decision makers easy and flexible access to (1) relationally structured data bases that contain the current state of the line (status of a machine, the location of a lot, the due date of a lot, the availability of an operator); (2) knowledge bases that contain such information as how to characterize a transaction (is it a lot movement, a change in the status of a machine, or change in an order), characterizing the lot type (is it a test lot from the lab, a test lot from manufacturing engineering,, an express lot for an important order, etc.), the setup required for a lot, setup time, rework requirements, test requirements, alert conditions, product routing, throughput rates, preferred tools, operator training, operator schedules, average down time for a machine, and how to calculate elapsed time. (3) Models that estimate how far ahead or behind schedule a lot is and the relative priority status of a lot, identify lots with the same setup requirements, determine daily output and/or range goals, establish global flow control levels (protective WIP, recommended output from a work cell for the day, and so forth) to guide production and avoid local optimization to the detriment of the global system, and assess the impact of machine dedication; and (4) heuristics to integrate the data, knowledge, and models to identify opportunities. An overview of the LMS structure is presented in Figure 14.18 (Overview of LMS Structure).

Appendix 5: Matching Assets with Demand

The heart of the production planning type activity in the tactical and operational decision tiers involves matching assets with demand (MAD). MAD refers to aligning assets with demand in an intelligent manner for a variety of purposes within the supply chain management (SCM) process. The alignment or match occurs across multiple manufacturing facilities within the boundaries established by the manufacturing specifications and process flows and business policies. Assets include, but are not limited to, starts, work in progress (WIP), inventory, purchases, and capacity (manufacturing equipment and manpower). Demands include, but are not limited to, firm orders, forecasted orders, and inventory buffer. The matching must take into account manufacturing or production specifications and business guidelines. Manufacturing specifications and process flows include, but are not limited to, build options, bill of material (BOM), yields, cycle times, anticipated date a unit of WIP will complete a certain stage of manufacturing (called a receipt date), capacity consumed, substi-

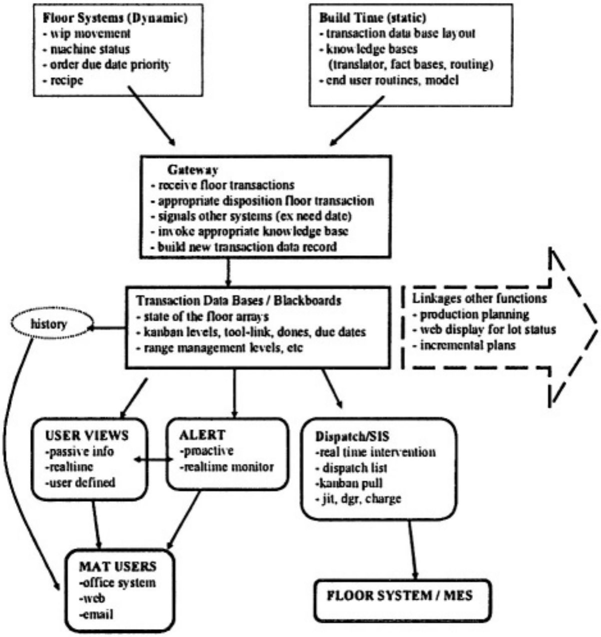

Floor Systems (Dynamic)
- wip movement
- machine status
- order due date priority
- recipe
Build Time (static)
- transaction data base layout
- knowledge bases
(translator, fact bases, routing)
- end user routines, model
Gateway
- receive floor transactions
- appropriate disposition floor transaction
- signals other systems (ex need date)
- invoke appropriate knowledge base
- build new transaction data record
Transaction Data Bases / Blackboards
- state of the floor arrays
- kanban levels, tool-link, dones, due dates
- range management levels, etc
history
Linkages other functions
- production planning
- web display for lot status
- incremental plans
USER VIEWS
-passive info
-realtime
-user defined
ALERT
-proactive
-realtime monitor
Dispatch/SIS
-real time intervention
- dispatch list
-kanban pull
- jit, dgr, charge
MAT USERS
-office system
-web
-email
FLOOR SYSTEM / MES

Figure 14.18. Overview of LMS Structure.

tutability of one part for another (substitution), the determination of the actual part type after testing (called binning or sorting), and shipping times. Business guidelines include, but are not limited to, frozen zones (no change can be made on supplies requested), demand priorities, priority tradeoffs, preferred suppliers, and inventory policy. Many of the manufacturing specification and business guideline values will change often during the planning horizon (called date effectivity).

References

Acock, M., and Zelmel, R. 1986, ".DISPATCHER: AI software for Automated Material Handling Systems" Proceedings: ULTRATECH: Manufacturing Automation Protocol, Artificial Intelligence in Manufacturing, Automated Guided Vehicles, Finstrat, Long Beach, CA /22 - 25/86, pp. 2139 - 2146.

Bitran, G. and D. Tirupati 1989, "Tradeoff Curves, Targeting and Balancing in Manufacturing Networks," Operations Research, Vol. 37, No. 4., pp. 547-564.

Bobrow, D. 1991, "AAAI-90 Presidential Address: Dimensions of Interaction," AI Magazine, Fall 1991, Vol. 12, No. 3, pp. 64-80.

Brown, J., Pakin, S. and Polivka, R. 1988, "Advanced Prototyping Language at a Glance," Prentice Hall, New Jersey.

"Business and the Internet Survey: You'll Never Walk Alone," 1999, Economist, June 26^{th}, pp. 11, 12, 17, 20, 21.

Chen, H., Harrison, M., Mandelbaum, A., Ackere, A., and Wein, L. 1988, "Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication," Operations Research, Vol. 36, No. 2, pp. 202-215.

Epp, H., Kalin, M., and Miller, D. 1989, "An interactive adaptive real-time scheduler for steel making," Proceedings of the Third International Conference: Expert Systems and the Leading Edge in Production and Operations Management edited by K. Karwan and J. Swiegart, Management Science Department, University of South Carolina, Columbia, South Carolina 29208, pp. 495-503.

Feigenbaum, E. et al., 1988, The Rise of the Expert Company: How Visionary Companies are using Expert Systems to Make Huge Profits, Times Books, ISBN 0-8129-1731-6, pp. 55-65.

Fordyce, K. and Sullivan, G. 1986, "Decision Simulation - One Outcome of Combining Expert Systems and Decision Support," in Artificial Intelligence in Economics and Management, edited by L. Pau, North Holland, New York.

Fordyce, K. Sullivan, G. 1999, "Supply Chain Management, Decision Technology, and e-business Information Technology at IBM Microelectronics," MicroNews, a publication of the IBM Microelectronics Division, Fourth Quarter 1999, Vol. 5, No. 4, pp. 18-21, `chips.ibm.com/micronews/vol5_no4/fordyce.html`

Fordyce, K. 1998, "Matching Assets with Demand Engines for PROFIT and Supply Chain Management," MicroNews, a publication of the IBM Microelectronics Division, Third Quarter 1998, Vol. 4, No. 3.

Fordyce, K., Norden, P., and Sullivan, G. 1987, "Links between Operations Research and Expert Systems," Interfaces, Vol. 17, No. 4.

Fordyce, K. and Sullivan, G. 1994, "Logistics Management System (LMS): Integrating Decision Technologies for Dispatch Scheduling in Semiconductor Manufacturing," Chapter 17 in Intelligent Scheduling, edited by Mark Fox and Monte Zweben, Morgan Kaufman Publishers, pp. 473-516.

Fordyce, K. and Sullivan, G. 1990, "Cycle time versus Machine Utilization: Moving Along the Curve Versus Shifting the Curve," IBM, TR 21.1440.

Fowler, J. 1992, "Issues in Semiconductor Manufacturing Scheduling," NSF Workshop on Hierarchical Control for Real-Time Scheduling and Manufacturing Systems, October 16-18, Lincoln New Hampshire sponsored by the National Science Foundation, edited by M. Caramanis (Boston University), S. Gershwin (MIT), and P. Vakili (Boston University).

Fromm, H. 1992, "Some Remarks on Cycle Time, Variability, Zero Inventories, and Costs in MicroElectronics Manufacturing Lines," IBM, German Manufacturing Technology Center, Sindelfingen, TR 28.167.

Galbraith, J. 1973, Designing Complex Organizations, Addison-Wesley, Reading, Massachusetts.

Graves, Meal, Stefek, and Zeghmi 1983, "Scheduling of Re-entrant Flow Shops," Journal of Operations Management, Vol. 3, pp. 197-203.

Hubns, M., Singh, P. (editors) 1998, Reading in Agents, Morgan Kaufman Publishers, San Franciso, ISBN 1-55860-495-2.

Jain, S., Barber, K. and Osterfeld, D. 1990, "Expert Simulation for On-line Scheduling," Communications of the ACM, Vol. 33, No. 10, pp. 54-62.

Keen, P. 1976, "Interactive Computer Systems for Managers: A Modest Proposal," Sloan Management Review, Vol 18, No. 1, Fall 1976, pp. 1-18.

Keen, P. 1980, "Decision Support Systems: Translating Analytical Techniques into Useful Tools," Sloan Management Review, Vol. 21, No. 3, Fall 19776, pp. 33-44.

Kempf, K. 1989, "Manufacturing Scheduling: Intelligently Combining Existing Methods," in Working Notes of AAAI AI in Manufacturing

Symposium, M. Fox editor, AAAI, 445 Burgess Drive Menlo Park, CA 94025-3496.

Leonovich, G. 1994, "An Approach for Optimizing WIP/Cycle Time/Output in a Semiconductor Fabricator," IBM Microelectronics Division, Essex Junction, VT 05452.

Lyon, P., Milne, R., Orzell, R., and Rice, R. 2001, "Matching Assets with Demand in Supply Chain Management at IBM Microelectronics," Interfaces, Vol. 31, No. 1, pp. 24-41.

Miller, D. 1990, "Simulation of a Semiconductor Manufacturing Line," Communications of the ACM, Vol. 33, No. 10, pp. 98-108.

Minnich, H. and Bula, H. 1985, "Closeup: Manufacturing control system at IBM tracks product movement through semiconductor line," Industrial Engineering, Vol. 17, No. 11 (November), pp. 82-90.

More, D., Pachinek, N., and Wong, D., "JAVA Based Mobile Agents," ACM Communications, March 1999, Vol. 42, No. 3.

Nilsson, N. 1998, Artifiical Intelligence: A New Synthesis, Morgan Kaufman Publishers, San Franciso, ISBN 1-55860-467-2.

Savell, D., Perez, R., and Koh, S. 1989, "Scheduling semiconductor wafer production: an expert system implementation," IEEE Intelligent Systesm, Vol. 4, No. 3, pp. 9-15.

Spearman, M. and Zazanis, M. 1992, "Push and Pull Production Systems - Issues and Comparisons," Operations Research, Vol. 40, No. 3, pp. 521-532.

Zweben, M. and Fox, M. (editors) 1994, Intelligent Scheduling, Morgan Kaufman Publishers, Menlo Park, Ca.

Chapter 15

CASE STUDIES: SUPPLY CHAIN OPTIMIZATION MODELS IN A CHEMICAL COMPANY

Young M. Lee
IBM T.J. Watson Research Center
Route 134, P.O. Box 218
Yorktown Heights, New York 10598
ymlee@us.ibm.com

E. Jack Chen
BASF Corporation
3000 Continental Drive - North
Mount Olive, New Jersey 07828
chenej@basf.com

Abstract In this chapter, we give a short overview of the supply chain management models that have been used in the past few years by one of the largest international chemical companies. These models have made significant impacts in improving strategic, tactical, and operational supply chain processes. Then, we describe three supply chain optimization model case studies in detail: distribution network optimization; capacity requirement planning; and Web-based production planning/scheduling. For each case study, we describe motivation for developing the supply chain optimization model, requirements, modeling methods, deployment and business impact of the model. Using these case studies we intend to share our lessons learned, and address supply chain management issues that are especially relevant to chemical industry. These models utilize mathematical programming, discrete-event simulation, and Web-enabling technologies.

1. Introduction

New business models and efficient management of supply chain are becoming critical success factors in today's highly dynamic and competitive business environment, driven by rapid advances in the information technologies and operations research methods (Geoffrion and Power, 1995). The goal of supply chain management (SCM) is to procure raw materials, manufacture products, and deliver the products to customers at desirable price and service. SCM requires coordination of the flow of products, services, and information among supply-chain entities, such as suppliers, manufacturers, distributors and customers (Keskinocak and Tayur, 2001). Many companies are using enterprise resource planning (ERP) tools to improve or optimize their supply chain. However, ERP systems often produce unrealistic production scenarios that result in excess inventories, sub-optimal utilization of resources and ultimately poor customer service (Hsiang, 2000). Therefore, it is necessary to model and optimize supply chain even if ERP systems are in place.

Modeling and optimizing supply chain management is much more affordable now due to relatively inexpensive computer hardware and abundant availability of supply chain modeling tools. The popularity of SCM tools is partly due to the advancement of the Internet, which allows easy access to such tools by supply chain decision makers. The Internet also facilitates supply chain coordination and collaboration with the suppliers and customers (Lee, 2002).

The study and work described in this chapter are based on supply chain management modeling activities of a large, international chemical manufacturing company. The company has been developing and using a wide range of supply chain management tools to better assess, analyze, and improve their supply chain. As typical supply chain characteristics of chemical industry, the company mainly produces functional products, which are defined as ones that have long product lifecycle and stable demand, with a relatively stable manufacturing process. Most manufacturing processes are continuous processes which require high initial capital expenditure to setup, and run as built-to-forecast processes. Production rates have a minimum and maximum range for physical reason as well as economic reason. Profit margin is relatively low; therefore, economies of scale are very important. The life cycles of chemical products are usually long.

For customer demand, the company has been using demand forecasting tools to predict future customer demand that uses historic patterns and anticipated business changes. Chemical companies mostly produce functional products, which tend to have more predictable demand than the innovative products such as high-end computers (Fisher, 1997; Lee, 2002). Nevertheless, the customer orders placed for manufacturing can exhibit significant fluctuations due to the "bullwhip effect" (Lee et al., 1997). Accurate forecasting, especially for individual product and longer time horizon, is very difficult because customer demands depend on many dynamic factors, such as economic, social, behavioral factors, and unexpected events. However, customer demand is often the main input for many strategic, tactical, and operational SCM tools, such as distribution planing, transportation planning, manufacturing planning, and capacity planning. Therefore, it is important to forecast customer demands as accurately as possible. Typically, demand forecast is done in aggregated product level and shorter time horizon to ensure that the forecast is reasonably accurate and meaningful. These demand forecast models are often based on time-series modeling.

The company has been developing and using many distribution network optimization models for finished goods distribution using tools, such as SAILS (Strategic Analysis and Integrated Logistics Systems, Insight, Inc.) and MIMI (Manager for Interactive Modeling Interfaces, AspenTech, Inc.). Some of these models are for packaged goods, which are usually shipped by trucks and stored in warehouses. For packaged goods, since different products can be shipped together and stored together, there are opportunities for consolidations of storage and transportation. Some other models are for bulk liquid products, which are carried by tank rail cars or tank trucks and stored in storage tanks or split into smaller volume in transloading (rail to truck) facility. Since bulk liquid products cannot be mixed together for transportation or for storage, bulk liquid network models are mainly used to identify optimal facility locations. Some of the distribution network models focus on global level and covers entire business region such as the NAFTA (North America Free Trade Agreement) region, and some other models focus on individual businesses, describing more detailed distribution process of a business. The output of the network model in one level is often used as input for a model in another level. These distribution network optimization models are supplemented by inventory analysis tools; for example, discrete-event simulation models that analyze dynamic effects of various replenishment policies and outbound (customer) shipment patterns including seasonal effect of sales. These simulation tools generate dynamic inventory profile at distribution facilities, which is critical in deciding the size of storage tanks and warehouse space. Distribution network models were developed using the Mixed-Integer Linear Programming (MILP) methods.

The company has also been developing and using production planning and scheduling optimization models. Production planning models usually focus on a single manufacturing plant or several manufacturing plants that produce common products, compute the optimal production amount of certain products in each production line for each time period, usually in weekly or monthly buckets. The models take into consideration demand, production capacity, storage capacity, and raw material availability. Finite scheduling models focus on daily or hourly sequencing of manufacturing equipment and other resources, and try to minimize the Work-In-Process (WIP) and inventory level. Production planning and scheduling models are especially important when manufacturing plants are running at full or almost full capacity. In chemical industry, many manufacturing processes are continuous; therefore, well-managed product changeover is very important in minimizing the interruption of the manufacturing process, which is time-consuming and costly. Production planning models are usually built using MILP methods, and the finite scheduling models are built using the combinations of MILP and heuristics.

Capacity planning is another area in which the company has been actively developing and using decision support tools. For examples, discrete-event simulation models have been used to determine the size of the railcar fleet that are used in transporting bulk materials from manufacturing plant to storage site and eventually to customers. We considered factors, such as transit time, customer dwell time (the length of time that the railcar stays in customer's premise), loading, unloading, maintenance and other activities, and identified the optimal number of railcars to ship products to customers on time for each business group as well as for the whole corporation. We also developed simulation models to determine optimal production and storage capacities.

In transportation area, we have been using a shipment consolidation model to optimally consolidate less-than-truck-load (LTL) shipments into a truck-load (TL) shipment and to compute optimal routes that minimize the total transportation time.

In the following chapters we describe three case studies of supply chain optimization models mentioned above, and discuss important issues in developing the models, implementing the solutions and the benefits. The three case studies are a finished good distribution network optimization model, a storage capacity requirement planning simulation model, and a Web-based production planning/scheduling optimization model.

2. Case Study 1: Finished Goods Distribution Network Optimization Model

The chemical company has been growing rapidly in the past few years through various acquisitions and divestitures. As seen as a common business practice in this dynamics business environment, the company has been constantly looking into adding businesses that improve its business portfolio, and divesting businesses that are not part of its core businesses. When a business is acquired, its distribution network is also inherited into the corporate network. Similarly, when a business is sold, a slice of the corporate network is removed. Over time the corporate distribution network became inefficient, and consisted of many independent and fragmented networks. The company wanted to assess the current distribution network, identify opportunity for network consolidation and improvement, and to implement the new optimal solution to realize the benefit as soon as possible. The company also wanted to put in place a network analysis tool that can be used periodically to study network as the network evolves with business.

The assessment study of the company's current network clearly showed that the network had several inefficiencies. One of them was operating too many distribution centers (DCs); more than 100 DCs were being used in the NAFTA region at the time of this study. When so many distribution centers are used, the amount of products stored and shipped from each DC is relatively small. Moreover, there were many small shipments that originate from each DC. The relatively small storage and transportation volumes make it difficult to obtain volume-discounted warehousing and transportation rate from service providers. Another inefficiency was customer assignment; some customers were serviced by DCs that are unreasonably far away. In certain cases, a DC located in the west coast of the U.S. was shipping products to customers in the east coast and vise versa. This was partly because individual business unit within the corporation was using only its own distribution network without utilizing other facilities available for the whole corporation. When the customer assignment is not efficient, the outbound shipment (customer-bound) lanes tend to be unreasonably long, thus increasing the transportation time and costs. Moreover, each DC needs to have certain level of safety stock to accommodate uncertain demands and unforeseen production problems (Brown et al., 2002). Long distance transportation to customers also requires relatively high amount of safety stocks in warehouses. Therefore, having both too many DCs and long transportation lanes contributeed to the high level of overall safety stock, and caused relatively large inventory holding costs. Thus, it was evident that the distribution network needed to be optimized.

The primary objectives of the network optimization project were to identify the optimal number and locations of strategic DCs, and to optimally assign customers to appropriate DCs to lower overall distribution costs and improve customer services. By consolidating DCs, the economy of scale can also be realized. With larger amount of products stored and shipped from each DC, we can have leverage for negotiating better storage and handling costs with warehouse service providers. With smaller number of transportation lanes and larger volume in each lane, we can also have leverage for negotiating better transportation costs with transportation service providers. The shorter distance between DCs and customer will also improve the customer services. The overall safety stock requirement will go down, therefore, lowering inventory carrying costs. Also, working with a smaller number of service providers simplifies the whole distribution process. Operating smaller number of DCs also allows more opportunities to consolidate cross business LTL shipments into TL shipments, which is much less expensive; therefore reducing the overall transportation costs.

Distribution network usually degrades over time, similar to entropy in thermodynamics, which naturally moves toward a higher degree of disorder. Especially in today's dynamics business world, it is very likely that a business will change in many unexpected ways altering the supply chain substantially. Distribution networks have to be re-evaluated every few years and be re-optimized. Another objective of the project was to make available a strategic distribution network optimization model that can be re-used periodically with simple changes of data and constraints, and to identify opportunities for network improvement.

2.1 Model

A very large scale MILP model was developed to model and optimize the finished goods distribution network. The model consists of roughly 40 manufacturing sites, 100 candidate DCs, 300 demand regions, 100 aggregated product groups, and million shipment transactions per year.

In the past twenty years, many studied have been done in modeling and solving complex distribution network problems (Geoffrion and Graves, 1974; Bradley et al., 1977; Brown and McBride, 1984). However, it is very difficult to develop a comprehensive, global distribution network model that would optimize over all aspects of businesses (Sweeney et al., 1997) in a large company because the model size would be too large to manage. Therefore, we modeled the distribution network in two levels; global and business level. In the global level, we modeled the entire business, e.g., the NAFTA region, and this model was primarily used for identifying optimal number and locations of strategically important distribution facilities. In this global level model, we used a higher level of product aggregation and simpler transportation rate structure to focus on the corporate-wide network rather than details of each business group. The solution from the global level model was passed onto the business level models. In business group level modeling, we focused on individual businesses, modeling many details of a distribution network of one business at a time. The business level models were primarily used for computing customer assignments and detailed cost calculations, which are needed for implementing the network solution. Individual business network models contained various business specific constraints, such as service level requirements, and used much less product aggregation and more accurate, lane by lane, transportation rate structures.

Transportation rates for replenishment flows (plants to DCs) were assumed to be all TL shipments, and we used a customized tariff for the company. Transportation rates for direct shipments (plants to customers) and outbound shipments (DCs to customers) consisted of TL and LTL shipments, and were computed as weighted average rate using the historic profiles of shipments and TL and LTL tariffs, which are company specific tariffs and are dependent on weight breaks. The historic shipment profile were computed by analyzing one year's transaction data from corporate ERP system.

The distribution network optimization model is a strategic tool that helps decide the optimal network structure by generating aggregated information, such as annual throughput at each facilities, annual transportation costs, and average service levels. However, the model is not adequate to address the dynamic effects, such as shipment size, frequencies of replenishment, outbound flow on the network, and seasonal demand fluctuations (Cheung et al., 2001). The dynamic effects are critical in analyzing dynamic profile of inventory and determining the safety stock, DC capacity requirement and inventory policies. We developed a discrete-event simulation model using eM-Plant (Tecnomatix, Inc) to supplement the network optimization model by analyzing dynamic profile of inventory. The simulation model was used to analyze various replenishment policies and outbound shipment patterns to determine the DC capacity requirements.

2.2 Results

The distribution network optimization models were successfully optimized for many business scenarios. A distribution network model is a rough description of a real distribution process with many assumptions and approximations. In business world, obtaining the global optimization solution and taking it blindly as the optimal business solution is risky. The distribution network model deals with many uncertain data, such as future customer demand, warehousing costs, transportation rates, and efficiency of managing distribution network. Therefore, finding the mathematically optimal solution is not as significant as finding insights and a set of solutions that can be rationalized and implemented easily. We experienced, in certain cases, that it takes many hours of computation just to improve a good solution by a fraction of 1%, which is much smaller than the error of data and is rather insignificant. We treated the distribution network optimization model as a decision support tool, which provided useful information to business decision makers so that a good decision is made and understood. The network solution we presented to the decision makers was a set of optimal network scenarios that clearly explained the tradeoffs among important components of distribution network such as distribution costs and customer services.

Our distribution network study indicated that the optimal number of DCs is between 6 and 10, with more than 95% of products reaching customers within next day delivery service (within 450 miles from a DC in our case). This result was also intuitive because 6 to 10 circles with radius of 450 miles should be sufficient to cover all the customers regions of the continental U.S. and Canada. The savings from the optimized network was several million dollars in transportation and warehousing costs, which was about 10% improvement of the network. The customer service improvement was about 35%, with most of customer serviced in next day delivery. The optimal distribution network was reviewed with all the business groups in the company, and was approved. Another benefit of the modeling was the fact that during the modeling process the organization has obtained much better understanding of the distribution network as well as its business.

There are many challenges in modeling distribution network, especially for large company that consists of many business units with their own business goals and needs. One of the challenges is obtaining the data required for the model. The optimization computation is based on data on model parameters, such as shipment transaction data, transportation costs, DC variable costs (handling and storing), and DC fixed cost. If the data were not accurate, the optimal results would be wrong, too. The company has been in the process of standardizing ERP systems when the model was being built. Therefore, there were more than one data sources; some businesses had their data in one ERP system, and others had data in another. Each ERP system had somewhat different data elements, format, unit of measure and naming convention, and it was time-consuming to unify the data from multiple sources into a consistent form.

Another difficulty of the modeling was to have all the business groups to participate in the modeling activity. It was extremely important that all the business groups provided the necessary data for the modeling, and validated the network solutions. The primary objective of the distribution network optimization was to consolidate distribution activities of all the business groups into a corporate-wide strategic network. The benefit was geared toward the corporate-wide optimization, not individual business optimization. Majority of business groups would benefit by participating in the optimization; however, a few business groups may not save money and even lose money as a consequence of implementation of new network solution. It was very difficult to convince those business groups to sacrifice for the benefit of the whole corporation since business leaders are compensated by the performance of their own businesses not by that of the overall corporation. And, there is always resistance to change. Implementation of the network solution involves changes in the business processes, and it is not a painless task. Moreover, there is often lack of trust on mathematical solution. Business leaders often feel that supply chain management is too complicated to model mathematically.

Modeling distribution network is relatively easier than implementing the network solutions. The implementation includes shutting down some DCs, which involves termination of employments. It also includes opening new DCs, which requires evaluating and selecting DC service providers from several candidates or even building private DCs. Requirements for new DC are computed from the network optimization model, and they include annual throughput, expected turns of products, special storage requirements such as for flammable and refrigerated products, frequencies and sizes of replenishment, and outbound shipments etc. Implementing the network also involves selecting new transportation service providers and canceling existing agreements. The transportation service requirements are transportation lanes, shipment profiles (frequencies and sizes), etc. The requirements for DCs and transportation are communicated to service providers with RFP (request for proposal). Once the proposals are received, they are carefully evaluated, and one that promises the least cost and the best service is selected. The network optimization models are critical not only for designing the network but also for generating the essential information required for the implementation of the optimal network.

3. Case Study 2: Web-Based Production Planning/Scheduling Optimization Model

One of the polymer manufacturing plants in the chemical company was facing problems of production capacity and production planning flexibility. Customer demand has been increasing, and due to the dynamics of economics it was more difficult to predict future customer demand for the future. The manufacturing process is a continuous process, and the process has to be interrupted often to switch over from producing one product to another, and each interruption idled the production for two weeks. There was one production planner at the plant who has to continuously communicate with a product manager at the business headquarter, and it took a few days for the planner to generate a production plan based on the demand input from the product manager. Quite often, a production plan has to be modified to accommodate changing customer demand, and it also took a few days to change the plan. The planner has been using a rather old planning spreadsheet to display the input data and to generate planning report. The company called for a better tool that can improve the quality of production plan, reduce the planning time and has the flexibility of rapidly modifying production on-demand. It was also important that the tool is used both by production planner at the plant and a product manager at the headquarter. We developed a MILP-based production planning optimization model that runs on the Web for this problem.

The plant has multiple production lines and produces hundreds of millions of pounds of a polymer per year. Each production line is a continuous process that operates 24 hours a day, 365 days a year except during maintenance periods when the production is interrupted for a few days. The raw materials are brought into the facility, usually by pipeline, and are fed continuously into the process. The plant produces multiple product grades with different physical properties. However, the product changeover cannot be done easily. When the plant switches from one product to another in a production line, the line will produce an off-grade product for a few days to weeks until the process reaches a steady state and produced a product with the desired specifications. Off-grade products can be sold but at a much lower price than products that meet the specifications, thus minimizing the number of product changeovers is important. Furthermore, the plant is operated at full capacity; therefore, it must have a well-planned product-changeover schedule to maintain overall production levels. Each production line can produce only certain product grades, and is constrained by a minimum and maximum production rate. The permanent storage tanks for the finished products have a limited space and temporary storage space is costly to use. Each production line has a minimum length for production campaigns during which product changeovers are not permitted. The goal of planning production is to compute a production plan that minimizes the inventory holding costs and product changeover costs while satisfying all the customer demands for finished goods and other process constraints.

The Internet greatly facilitates the deployment of highly interactive applications. With the Internet, it is now possible to deliver computational services that were once available only to those employees with specialized computer training and access to special computers and tools. The Internet-based computational applications can be accessed from virtually any Web browser on any computer anywhere in the world at any time to perform complex computational tasks. Optimization is one such computational tool that can provide lots of benefits when it is available on the Internet. Optimization has been used widely in industry for solving complex business problems. The company has been using MILP to optimize distribution networks, production planning, and scheduling. However, until recently, the users of such optimization applications had to have powerful computers with special optimizing engines and other data interface utilities. Many of the optimization models used in the company have been standalone applications and have lacked standard interfaces with other enterprise applications, such as data warehouse (DW) or ERP systems. Therefore, it has been difficult to deploy such optimization tools to multiple users throughout the company. Also, communicating optimization results among users have not been easy.

With the optimization tools on the Web, virtually anyone within the allowed community on the Internet or Intranet can access the complex optimization tools without any special hardware or software. It is now easy to make optimization technology available to many people. The optimization engine can reside on only one powerful server with enough computing resources (or in rare cases, a few servers). Moreover, a network of computers can serve as a parallel and distributed processing server environment for solving computationally intensive problems. Communicating the optimization results among the users, especially among business managers, engineers, and production planners, is easy with the Web-based tool, because the optimization results are stored in a centralized server and can be accessed by and presented to the users through a flexible and powerful medium, such as Hyper Text Markup Language (HTML). Furthermore, maintaining a Web-based optimization tool is much easier than maintaining traditional optimization tools installed individually on each user's computer. One can modify or enhance a Web-based optimization tool on one server, and all the users can access the change immediately. Supply Chain Optimization tools are particularly well suited to Internet innovations.

Therefore, we designed, implemented, and deployed an interactive Web-based optimization tool for this production planning optimization problem. The framework we developed is general and modular, and it can be used for developing similar tools for other businesses with the corporation. This tool permits users to change the objective functions and constraints of optimization models using a Web-browser and to run optimization and view the results in HTML pages. The users do not need to use FTP or TELNET protocols. In our framework, the input and output presentations are dynamically generated from a JSP (Java Server Page) that resides on a Web server. The Web is a client-server application; the client is a local computer and the server is a remote host (computer). The input data are taken from the clients and passed to the application (optimization) server, where an optimization model is executed remotely. Typically, the server is a powerful, high-end computer. After the optimization is complete, the results are passed from the server back to the client computers in the form of a standard HTML document, which users can view on the browser. The client computer could be any computer with a Web browser and an Internet access.

3.1 Model

We formulated the production planning problem as a MILP model, and used XPRESS-MP to model the problem and to optimize the model. Lee and Chen, 2002, describe the details of the mathematical formulation for this model.

Java technology has revolutionized computer use, and many Web-based applications are being developed in Java. However, most optimization modeling and optimization packages, such as ILOG OPL (Optimization Programming Language) and XPRESS-MP (by Dash Associates), are based on C and other traditional programming languages. Thus, it was impossible to call those optimization subroutines from Java directly until recently. Fortunately, Java provides the Java Native Interface (JNI), which allows Java to interface with other popular programming languages, such as C or C++, Visual Basic, and Fortran, which can be interfaced with most optimization packages. We developed a framework for calling optimization subroutines from Java via JNI with Web browsers.

Figure 15.1 shows a three-tiered architecture for web-enabled optimization tools. The first tier on the client side processes the input data and presents the output data. The second tier is the Web-server, which manages the server-side processing and communicates with third tier servers such as the database server and the application server. The JSP and Java programs, the database system, and the optimization engine can all run on one server. However, because of the security and performance reasons, it is better to run them in separate servers, for example, on a Web server, a database server and an application (optimization) server. Users can use a Web browser in the client computer to edit input data, invoke execution of the optimization, and receives results via Web pages. The main implementations and processing tasks are carried out on the server side. This framework provides the flexibility of a programming language in a production environment, and developers can customize connections among models, data sources, and user interfaces.

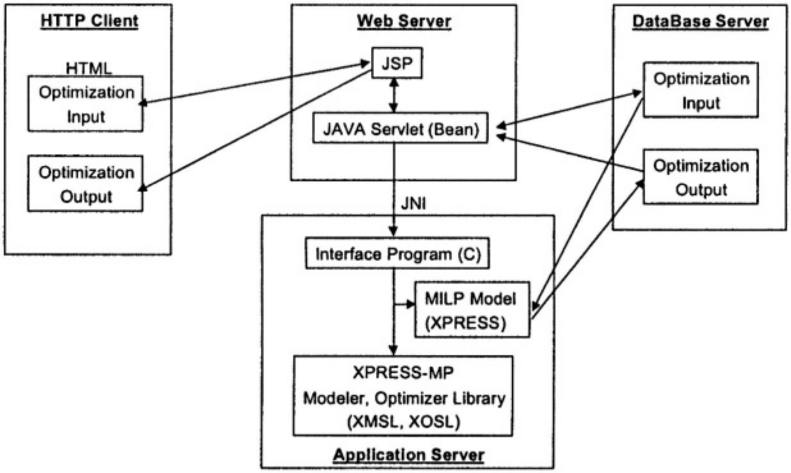


Figure 15.1. Architecture of the Web-based optimization

Some optimization-service Web sites deliver Web-based optimization by providing FTP utilities; users upload their local model and data files to the host computers and remotely invoke the execution of the optimization model. Some other optimization-service Web sites provide text boxes that allow users to edit their model and data files in the server. These implementations are fairly easy and straightforward. However, they require users to know quite a bit about the optimization model and the optimization engine to use the tools.

In chemical industry, however, users of optimization tools typically are not people trained in mathematics or optimization. Therefore, to be useful, optimization tools must be easy to use. In our implementation, we provided a user-friendly interface to allow people with little knowledge of mathematical modeling to easily operate the optimization model. Users of the models control their optimization goals, such as minimizing cost or maximizing profits, and constraints by changing the parameters in the user interface screen; however, they often don't need to understand the mathematical models and solver to use the model.

The optimization process starts when the users make an HTTP request for a JSP from a client computer. The JSP technology enables rapid development of dynamic Web applications that are platform independent. JSP separates the user interface from content generation, making it possible to change the overall layout of the Web page without altering the underlying dynamic contents. Pekowsky, 2000, describes how JSP works with HTML in detail. The JSP program makes a connection to optimization input files via a Java Servlet (JavaBean) and dynamically displays the names of the optimization input data tables associated with the model.

Using the input-file-selection page, the user selects one or more tables from the files, for example, demand forecast; then a JSP generates and displays an HTML page with those data. The data are displayed in a tabulated format; values that can be updated are displayed in text cells. Users can change the data by typing over the displayed data in the text cell on the browser. The modified data updates the files through the JSP and a servlet. Most solvers provide the option to decouple the model and data files. The high-level algebraic formulations describe optimization models in concise, symbolic formats, and an accompanying data file specifies the model instance to be solved. However, different solvers may require different formats for the model and data files. A servlet will translate data from the browser into solver-specific formats. Moreover, some solvers provide the option to read in data from spreadsheets and databases based on the SQL queries in the optimization model. If solvers do not have the option to interface with spreadsheets and databases directly, one can write a Java servlet to retrieve data from databases, translate data into solver-specific formats, and write them to data files. It is also possible to retrieve or update data from multiple databases using different access methods and protocols as long as they are available on the network.

Once the users update the input data, the JSP program calls a servlet that runs a JNI with a C-based program, which in turn runs the optimization engine with an optimization model. The C-based program will initiate a command instructing the optimization model to read in data from data files specified in the model and to bind the data with predefined variables before it initiates the optimization command. Thus, the model and data are completely independent. When the optimization is completed, the results are updated to the output files. A JSP program then dynamically displays the selected output report tables to the user's browser as HTML pages. The user can select any tables and view the optimization results. When the user selects the output file of interest from the drop-down menu; a JSP generates and displays an HTML page with those data. A downstream application uses the optimization results to generate daily production schedules, hourly raw material feed rates, and production reports, such as Gantt charts.

The framework will work with any optimization packages as long as they have C-based optimization library. Lee and Chen, 2002, detail the implementation of this architecture.

3.2 Results

The production-planning model we implemented has 1,452 variables (308 binary variables) and 1,113 equations. The optimization usually takes a few seconds on a Sun/Solaris workstation, running Netscape Enterprise Server. The response time depends on other factors as well, such as the state of the application server and the load of network traffic. Before we implemented the model, the production planner took several days to plan a production schedule. The short response time of the integrated model has allowed the production plant to adjust its production schedule quickly to accommodate any sudden market changes. The quality of production planning has also improved. The planning model has helped the production planners to reduce inventory and to improve the utilization of production and storage capacity. In addition, business managers in different locations are now able to view the results and make intelligent business decisions quickly. The maintenance and technical support of the model have become much easier too. We modify and enhance the tool in only one server, and all the users can access the change immediately.

One of the benefits of integrating the optimization engine with the Web is that we can easily implement a parallel and distributed processing capability into the infrastructure. So far, application of parallel processing has been limited. Until recently, parallel computers could be found only in research laboratories or large universities. Furthermore, system software to support large-scale distributed processing remains scarce (Luo et al., 2000). On the other hand, an inherent characteristic of the Web is its distributed processing nature. The Internet can emulate the parallel processing architecture of expensive parallel hardware, while the Internet protocol unifies diverse networking technologies and administrative domains. A network of several computers within the Intranet or Internet can collaborate to solve complex optimization problems, effectively utilizing computers that may be unused otherwise. In solving an MILP problem, for example, a main server can generate sub-optimization problems through the branch-and-bound method, and several computers within the network can optimize the sub problems in parallel while the main server orchestrates the overall optimization strategy. Commercial optimization software, such as XPRESS-MP, can support such parallel-processing architecture. Therefore, the marriage between Web-based optimization and parallel and distributed processing seems natural.

The Web-based optimization infrastructure we developed is a generic framework that can be applied to a variety of optimization applications. The JNI interface between the Java class and the C-Interface program is generic and can be customized easily for other Web-based computational applications.

Once an Operations Research tool becomes available through the Intranet or Internet, it can be further integrated with other enterprise applications. For example, the input parameters of an optimization model can be updated in databases regularly through some other enterprise applications. End users of the optimization tool do not need to modify the input data themselves. Moreover, the optimization results can be stored in databases for other applications.

4. Case Study 3: Storage Capacity Requirement Planning Simulation Model

The company was planning a major production capacity expansion of one of its major product plants. The upstream of the process, i.e. manufacturing, was already scaled up by a group of engineers, and the company wanted to make decisions on the downstream processes, i.e. storage, mixing, packaging and transportation. There were a number of existing silos for mixing and storage, and packaging equipment for the plant. Adding new equipment, especially silos, is very expensive. However, sufficient silo space is very important. A shortfall of the silo space will interrupt the manufacturing process because the process is continuous and the product coming out from the manufacturing end would not have any place to go. The shortfall also affects the transportation process thus affecting customer service level as well. The company called for an accurate analysis to decide whether, how many and what size of silo and packaging equipment are needed for the plant. We developed and used a discrete-event simulation model to analyze the problem. The simulation model helped us in determining capital equipment requirements and assessing alternative strategies for the logistics operations.

The plant produces three different grades of a dry chemical (denoted as A, B and C) at a specific production rate. These three different grades are produced in a continuous cycle with a fixed quantity for each grade. The product is transferred to a storage tank, from which it is distributed to another facility of further processing and packaging. A larger portion of products is sent to railcar for shipment, and the rest is sent to truck for shipment. Furthermore, the sequence of railcar and truck shipment is random and mixed. The capital outlay of such facilities is tremendous, and the designer needed a credible, valid, detailed model of operation.

There are several large volume silos available for the plant. However, only one silo can be used to receive the production outflow from the plant at any given time. The outflow from the silos cannot take place until the silo has completely filled. This is necessary because a batch number will be assigned to a particular silo load so that the source and quality of the product can be traced. Only one outflow from the silos can take place at any given time. Grade A of the product requires special blending and needs to be kept in the silos for at least twenty-four hours. There is one RailSilo used to load railcars, which has a loading capacity that is a multiplication of railcar load to avoid less-than-full-load railcars. The RailSilo cannot have flow-in and flow-out at the same time. There is one BagSilo used for the bagging process. The BagSilo has a smaller capacity than other silos, however, it can have flow-in and flow-out at the same time. The flow-out rates from all the silos are all fixed.

While bulk railcar shipments do not require special packaging, truck shipments need to be bagged first. The bagging-process will produce a certain volume of bagged products every few minutes. The bagging machine requires a few minutes of maintenance after processing a certain volume of the product. It takes a few minutes to change over between different grades of products. The bagging machine breaks down occasionally and needs to be stopped for repair. When trucks arrive at the plant, they are weighed at the weigh station, and the process take a few minutes. A fixed fraction of the arriving trucks are here to pick up our bagged product. The remaining fractions are here for other purposes. There is a fixed number of loading docks, and it takes a few minutes to load the truck, which also has a fixed capacity. Once the truck is loaded, it needs to be weighed again before it can leave the premises. Both the inbound and outbound trucks use the same weigh station. If there are more than one truck waiting for the weigh station, the order of trucks go to weigh station will be based on first come first served basis.

The main objective of the simulation study is to ensure the process configuration and capacity can support continuous outflow of the manufacturing plant, and to optimize the number and size of storage silos. There are several standard size silos under consideration. The decision is not only to select more silos with smaller size or fewer silos with larger size but also to optimize the combination of the number and size of silos. Moreover, the activities of bagging process, the activities of railcars and trucks, such as the inter-arrival time of railcars and trucks, are analyzed to ensure continuous material flows are maintained without interruption. Once the average inter-arrival time is determined, the size of railcar fleet can be calculated indirectly. The simulation model can help us not only to verify the feasibility of our configurations but also to search for the optimal configuration among several alternatives.

4.1 Model

Many production activities in chemical industry involve continuous material flows, such as liquid, gas or solid, and it is very costly to interrupt and restart the production process. Simulation is a useful tool to study dynamics in such processes in a simulated environment. Simulation models do not only provide quantitative information that can be used for decision-making but also increase the level of understanding of how the process works. Most models are used to simulate discrete events. Discrete-event simulation concerns the modeling of a system as it evolves over time by a representation in which the state variables change instantaneously at separate points in time (Law and Kelton, 2000) and has a commendably long and successful track record in the improvement of manufacturing process (Law and McComas, 1997).

Although in the chemical manufacturing plant, materials are mixed and transferred as continuous flow through a maze of tanks and pipes, we did not have to model the continuous components to effectively study throughput issues. We defined the product in a batch that uses various resources for a period of time simply by the amount of fluid being transferred and the rate of transfer. We used discrete-event simulation to model the continuous material flow in this plant. In many real world applications the behaviors of discrete event and continuous process are often interdependent. Note that several simulation packages have the capability to build hybrid discrete/continuous models. Some researchers have developed simulation models to analyze the hybrid nature of chemical manufacturing plant (Watson, 1997; Saraph, 2001). Some simulation issues in this area are conceptualizing production operations for simulation, discretization of continuous processes and building adequate level of detail in the models (Chen et al., 2002).

The output of the manufacturing plant is continuous at certain metric tons per minute. The transfer of a continuous flow from Silo X to Silo Y via Pipe S was simulated as a delay based on the amount being transferred and a fixed transfer rate. We discretized the continuous material flow to a fixed weight moving unit. The output was then converted as one unit every period. The weight per unit was initialized from a data table in the model. For example, if the output rate is 6 metric tons per hour, and the weight per discretized unit is 2 metric tons, then the output rate becomes 3 units per hour or one unit every 20 minutes. In general, with a smaller discretized unit weight, the simulation model can simulate the continuous material flow more accurately. The model can be regarded as continuous if the discretized unit weight is the weight of a grain of the product. However, it also complicates the simulation model because most discrete-event simulation software uses the next-event time advance mechanism for the simulation clock (Law and Kelton, 2000).

We wanted to build a simulation model that allowed us to analyze the logistics system adequately without modeling unnecessary details. We chose two metric tons per unit for our model because it is the smallest incremental weight that the product are bagged and processed in this logistics system. This discretized unit weight allowed us to analyze the system adequately without complicating the modeling of the bagging process. If there are two different packaging sizes, for example 2 and 5 metric tons, the 2 metric tons per unit discretization will complicate the implementation of the simulation model. In this case, we would use one metric ton, which is regarded as the smallest incremental weight per unit.

One of the purposes of the simulation analysis was to find out the minimum required number and size of storage silos; therefore, the outflow control from the plant always searched the available silos from left to right as the downstream station. Thus, excessive silos will not be used by the system. The I/O control between the main silos and RailSilo, BagSilo determined which main silos should have outflow and which downstream silos the material should flow to. The outflow of main silos was based on the "first available" rule. The flow-in time was recorded when the material in the silo was ready to flow out. For example, grade A product may be stored in Silo 1 before grade B product is stored in

Silo 2. But the flow out of grade A product cannot take place until the material has been processed in the silo for at least 24 hours. Therefore, the I/O control will select Silo 2 for outflow instead of Silo 1.

The bagged product was stored in the warehouse until a truck made a request. The warehouse was viewed as a sink of upstream stations, i.e., the warehouse had virtually unlimited storage capacity. However, the warehouse acted as a source for downstream stations. The material was stored in the warehouse until a truck was ready to be loaded. To reduce the warm-up period, we assumed that there is a certain volume of initial inventory in the warehouse.

One of the difficulties in developing this model was to simulate changes of the statuses of the silos. Once a silo was completely filled, there was no further inflow until the silo was completely emptied. The outflow of the silo became available immediately when the silo was filled, except grade A which needed to be kept for at least 24 hours. A complication arises because there is a lag between the outflow from the upstream station to the inflow of the downstream station. It will be too late to switch outflow from the upstream station when the receiving silo is completely filled, because the material in the pipeline will be lost. Thus, it was important to synchronize all the processes in this model. For example, the plant needed to send its outflow to other silos when the material in the pipeline filled the receiving silo completely. We accomplished the synchronization by making the material move instantaneously. As soon as the material left a station, it immediately appeared in its destination. The transfer time between stations was simulated after it reached its destination. The outflow control was embedded in the silo object, which can adjust the flow out rate. For example, if the current material flow is from Silo 2 to RailSilo, then one unit will be removed from Silo 2 every few minutes according to the flow out rate. The unit was added to the RailSilo as soon as it has been removed from upstream. This was possible because the capacity of the inflow rate of downstream was always greater than the upstream outflow rate.

The arrival of railcars and trucks were modeled as Poisson processes with mean inter-arrival time of a few hours. Previous experience indicates that the stochastic arrival process can be adequately simulated with the Poisson process, i.e. exponential inter-arrival, and the interval between break down and the time required to fix a machine can be simulated with a Weibull distribution (Law and Kelton, 2000).

The visualization of the simulation model was very useful for users to validate the model. Visualization was also critical in communicating the outcome of a simulation study to the management. Decision-makers often do not have the technical knowledge to understand the statistical outcome of a simulation run. But through the visualization, the managers were able to see the status of the silos and the flow of material. The process of building the simulation model also gave an opportunity for the plant personnel and upper management to better understand the logistic process.

4.2 Results

Sargent, 2000, described several methods to validate simulation models, such as animation, historical data validation, face validity, extreme condition tests, internal validity, and traces. To reduce uncertainty, we used historical data to build and drive the simulation model whenever possible. Face validity refers to asking people who are familiar with the process whether the model and its behavior are reasonable. We used their feedback to determine whether the logic in the conceptual model was correct. We validated the model through several extreme conditions, where the analytic solutions were attainable. The model output was then compared with the verified analytical results. For example, if we set up the simulation model to terminate in one month, we can verify whether all the material adds up. We can trace the material in certain states, such as the quantity shipped by railcar and truck, the quantity stored in different silos, the quantity processed by the bagging machine, etc. Accurate statistical analysis is central to the validity of any simulation project (Law and Kelton, 2000). Since we were simulating stochastic systems, we could not conclude our results with one simulation run. Internal validity refers to make several independent runs of the model to determine the stochastic variability. A high variability may indicate the system is sensitive to its input parameters, and the appropriateness of the simulation results needs to be investigated more closely.

The users agreed that our model was an accurate representation of the real system. To alleviate any concerns of the robustness of the results due to the random variations inherent in simulation, each scenario was run multiple times with different time horizons, one, two and three years. The modeling approach described above was used to evaluate various alternatives. Many of the alternatives were defined and modified only in the data tables. This flexibility allowed the user to read in data, run a scenario, and get results very quickly. No scenarios required modifications to the model itself. Moreover, when the modifications are necessary, the model can be easily and quickly changed due to the object-oriented design of the model. The results from the simulation provided a clear picture as to a best choice of planning.

Silo	Utilization
1	63.98%
2	63.91%
3	10.93%
4	0.00%

Table 15.1. Silo utilization statistics

Several scenarios with different numbers and different sizes of silos were used in our experiments. The scenario study provided valuable information, because the cost structure of the size of the silos was not linear. The optimal combination of the number and size of silos was determined with simulation of a pre-determined set. After several preliminary experiments, we determined that three mid-size silos are most cost effective and are able to support the continuous operation of the manufacturing plant. The followings are experimental results corresponding to the model. Since we hypothesized that three main silos will be enough to support continuous flow, we set up four silos in the model so that we will be able to verify our hypothesis. Of course, we can model the system with three silos and check whether an overflow occurred, however, we will not have the utilization information of the non-exist silo.

Table 15.1 lists the silo statistics for one particular replication when the model was simulated with one-year time frame. The report indicates that the fourth silo has not been used, thus, it is possible to remove the fourth silo without causing disruption of the production flow. The low utilization of Silo 3 is also very re-assuring. If the fourth silo has been used, it is an indication that three silos are not enough to support continuous flow. Table 15.2 lists the bagging machine statistics. The report shows the utilization of the bagging operation is quite low at 43.65%, i.e., it is idle 56.35% of the time. In this plant, the designer purposely built a high-capacity bagging machine to accommodate the anticipated future expansion of the production capacity. Furthermore, the bagging machine is relatively inexpensive to build and operate. The report also shows that only 2.85% of the simulation time was used in changeover between different grades of product, and 1.66% of the time was used in maintenance.

The changeover time between different products is different. For example, it may take 20 minutes to switch from bagging Grade A to Grade B and take 30 minutes to switch from bagging Grade B to Grade C. The changeover information is stored in data tables, therefore, the bagging process will be able to simulate multi-products without any modification. The simulation results also provided information regarding the

Bagging Machine	Percentage
Idle	56.35%
ChangeOver	2.85%
Maintenance	1.66%
Grade A	5.98%
Grade B	19.50%
Grade C	13.66%

Table 15.2. Bagging machine statistics

number of changeovers and the average time between changeovers. This information was important in determining the campaign volume.

5. Conclusions

Chemical industry has unique supply chain characteristics such as continuous and stable production processes, large changeover cost, handling of bulk material, high volume per SKU (Stock Keeping Unit), long life cycle of products, and low profit margin. In this chapter, we described an overview of practical supply chain management applications that have been used in a chemical company.

We also focused on three case studies and discussed the motivations of developing such tools, values that those tools have added to the company, issues that needed to be dealt with, and lessons learned. For the first case study, we described a large-scale MILP model that was developed to optimize distribution of finished goods. The optimization model consolidated distribution network that consisted of many independent and fragmented networks. The model generated substantial savings in distribution costs and drastically improved customer services. For the second case study, a generic computation framework for web-based optimization was described. The framework was developed using a server-side Java programming, and a practical production planning optimization model was successfully developed and deployed using the framework. The model improved the quality of production plan, flexibility of production plan change and accessibility of the tool. For the third case study, we described how we used a discrete-event simulation to model a logistic process and to determine capacity requirements of the storage and packaging facilities that allow a continuous production outflow and customer shipments. The simulation model reduced a capital expenditure substantially.

Acknowledgement

The authors thank the anonymous referee for valuable comments, which greatly improved the quality of this chapter.

Online References

Underlined terms in the paper indicate online references.

- INFORMS Resources (www.informs.org/Resources/Computer_Programs)
- Java (developer.java.sun.com/developer/onlineTraining)
- Neos Server for Optimization (www.mcs.anl.gov/neos)
- PaperSuite 2.0 (www.optamaze.com)
- Remote Interactive Optimization Testbed (riot.ieor.Berkeley.edu/riot)
- SAS Institute (www.sas.com/solutions/supplychain/demos/index.html)
- Servlet (java.sun.com/products/servlet/2.1/)
- Statit tool set (www.statware.com)

References

Bradley, G.H., Brown, G.G., and Graves, G. (1977). Design and implementation of large scale primal transshipment algorithms. *Management Science,* 24(1):1–34.

Brown, G., Keegan, J., Vigus, B., and Wood, K. (2002). The Kellogg Company optimizes production, inventory, and distribution. *Interfaces,* 31(6):1–15.

Brown, G. and McBride, R. (1984). Solving generalized networks. *Management Science,* 30(12):1497–1523.

Chen, E.J., Lee, Y.M., and Selikson, P.L. (2002). A simulation study of logistics activities in a chemical plant. *Journal of Simulation Practice and Theory,* 10(3-4):235–245.

Cheung, W., Leung, L., and Wong, Y.M. (2001). Strategic service network design for DHL Hong Kong. *Interfaces,* 31(4):1–14.

Fisher, M. (1997). What is the right supply chain for your product. *Harvard Business Review,* 75(2):105–116.

Geoffrion, A.M. and Graves, G.W. (1974). Multicommodity distribution system design by Bender decomposition. *Management Science,* 20(5):822–844.

Geoffrion, A.M. and Power, R. (1995). Twenty years of strategic distribution system design: An evolutionary perspective. *Interfaces,* 25(5):105–127.

Hsiang, T. (2000). The illusion of power. *OR/MS Today,* 28(1):34–36.

Keskinocak, P. and Tayur, S. (2001). Quantitative analysis for internet-enabled supply chains. *Interfaces,* 31(2):70–89.

Law, A.M. and Kelton, W.D. (2000). *Simulation Modeling and Analysis.* McGraw-Hill, New York, New York, 3rd edition.

Law, A.M. and McComas, M.G. (1997). Simulation of manufacturing systems. In *Proceedings of the 1997 Winter Simulation Conference, Atlanta, Georgia,* pages 711–717, New York, New York. IEEE Press.

Lee, H.L. (2002). Aligning supply chain strategies with product uncertainty. *California Management Review,* 44(3).

Lee, H.L., Padmanabhan, V., and Whang, S. (1997). The bullwhip effect in supply chains. *Sloan Management Review,* 38(3):93–102.

Lee, Y.M. and Chen, E.J. (2002). BASF uses a framework for developing web-based production-planning-optimization tools. *Interfaces,* 32(6):15–24.

Luo, Y.C., Chen, C.H., Yücesan, E., and Lee, I. (2000). Distributed web-based simulation optimization. In *Proceedings of the 2000 Winter Simulation Conference, Orlando, Florida,* pages 1785–1793, New York, New York. IEEE Press.

Pekowsky, L. (2000). *Java Server Pages.* Addison-Wesley, New York, New York.

Saraph, P.V. (2001). Simulating biotech manufacturing operations: issues and complexities. In *Proceedings of the 2001 Winter Simulation Conference, Washington, D.C.,* pages 530–534, New York, New York. IEEE Press.

Sargent, R.G. (2000). Verification, validation, and accreditation of simulation models. In *Proceedings of the 2000 Winter Simulation Conference, Orlando, Florida,* pages 50–59, New York, New York. IEEE Press.

Sweeney, D.J., Dill, F.A., Wegryn, G.W., Evans, J.R., Camm, J.D., and Chorman, T.E. (1997). Blending OR/MS, judgement, and GIS: Restructuring P&G's supply chain. *Interfaces,* 27(1):128–142.

Watson, E.F. (1997). An application of discrete-event simulation for batch-process chemical-plant design. *Interfaces,* 27(6):35–50.